RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Doctorat de l'Université de Toulouse

préparé à l'Université Toulouse - Jean Jaurès

De la manipulabilité à la méthodologie et vice-versa : un regard critique sur les pratiques expérimentales dans l'étude des représentations d'objets

---

Thèse présentée et soutenue, le 12 décembre 2024 par

Dimitrios PAISIOS

École doctorale
CLESCO - Comportement, Langage, Éducation, Socialisation, Cognition

Spécialité
Psychologie

Unité de recherche
CLLE - Unité Cognition, Langues, Langage, Ergonomie

Thèse dirigée par
Nathalie HUET et Elodie LABEYE

Composition du jury
M. Ludovic FERRAND, Président, Université Clermont-Auvergne
Mme Penny PEXMAN, Rapporteure, Western University
Mme Solène KALÉNINE, Rapporteure, Université de Lille
Mme Nathalie HUET, Directrice de thèse, Université Toulouse - Jean Jaurès
Mme Elodie LABEYE, Co-directrice de thèse, Université Toulouse - Jean Jaurès

From manipulability to methodology and back: a critical look into experimental practices in the study of object representations

Perhaps it is best not to ponder too deeply issues of method – let's get on with our work and all will turn out fine in the long run. However, I believe that ignoring issues of method exacts its price, sometimes one we can ill afford.

Alfonso Caramazza (p. 42, 1986)

Acknowledgements

So many people contributed in some way or another to this work and to my life in the past four years that it is impossible to mention them all here. Thank you all for being there and for making it such a rich experience.

Je dois mes remerciements les plus profonds à mes encadrantes, Elodie Labeye et Nathalie Huet. Voilà maintenant 7 ans que notre aventure passionnante a commencé! Même si je suis triste de voir ce chapitre de ma vie se clôturer, je n'aurais pas pu en espérer un plus beau. Je sais que l'expression ne te plaît pas beaucoup Elodie (désolé!), mais merci d'avoir été de véritables "mamans/grandes sœurs académiques". Vous avez été des piliers dans ma vie au cours de ces années et m'avez fait énormément grandir, autant scientifiquement qu'humainement. Merci de m'avoir tiré vers le haut et d'avoir pris soin de moi dans les moments difficiles. Merci pour votre présence, pour votre écoute, pour votre bienveillance, pour votre confiance et pour votre amitié. Merci aussi pour votre patience face à mes *petites* manies et à mon amour pour les deadlines, et pour avoir su cadrer mon esprit qui aime s'éparpiller tout en me laissant toute la liberté dont j'avais besoin. Un énorme merci à Elie Prudhomme également pour m'avoir accompagné de loin tout au long de cette thèse sur le plan technique. Tu m'as sauvé d'innombrables heures de travail et de frustration. Merci d'avoir consacré gratuitement autant de ton temps et de ton attention, notamment pour la forme de ce manuscrit et pour m'avoir permis d'accoucher de ce bébé quand je commençais à perdre espoir!

I would like to thank my two reviewers, Solène Kalénine and Penny Pexman, as well as my examiner, Ludovic Ferrand, for accepting to take part in the jury, for their time and for our stimulating discussions. Merci également aux membres de mon CSI, Ludovic Ferrand et Nabil Hathout, pour avoir accepté de suivre ce travail et pour leurs retours au cours de la thèse.

Je suis très reconnaissant d'avoir pu partager cette aventure avec les doctorants et les post-doctorants du laboratoire, en particulier l'équipe des représentants. Merci infiniment de

m'avoir toujours donné envie de venir au labo, pour votre amitié et pour votre soutien, mais surtout pour tous ces projets ambitieux que nous avons entrepris ensemble. Grâce à vous j'ai appris à préparer des olives, découvert les différentes saveurs des chips Brets, chanté Petit Papa Noël avec ma plus belle voix (hein Flo?), fait décrypter un message extraterrestre aux membres du labo... Mon seul regret est de n'avoir jamais eu le temps d'aller à cette mystérieuse boulangerie clandestine. Merci pour toutes ces expériences précieuses, restez frais!

Victoria, merci pour ton énergie inépuisable et d'avoir toujours ensoleillé notre bureau, d'avoir partagé la réputation de "finisseurs de bouteilles", pour tes merveilleuses anecdotes et d'avoir embêté les autres doctorants en faisant des bruits bizarres avec moi. Claire, désolé pour les bruits bizarres et merci pour cette merveilleuse expérience de voisinage (toc toc), pour m'avoir appris à dire UNO, pour toute ton aide sur le manuscrit, pour avoir été ma confidente et amie, et de ne pas m'en avoir trop voulu pour le poisson.

Un grand merci aux membres, au personnel et à la direction du laboratoire CLLE pour leur accueil chaleureux et pour le soutien émotionnel, scientifique, administratif et matériel tout au long de la thèse. Merci également aux membres permanents dont les initiatives et engagements aident à faire de ce laboratoire un espace aussi enrichissant et ouvert. Pierre, merci pour toute ton aide sur les expériences, de m'avoir aidé à faire la déco de CLOE et pour toutes nos discussions passionnantes.

Chers mousquetaires, Elohan, Maxime, Valentin et Quentin, merci d'être les personnes engagées et critiques que vous êtes, de m'avoir inspiré à toujours faire mieux, et de m'avoir fait faire la "danse turque" à tant de soirées. Quentin et Valentin, merci d'avoir accueilli cet étranger pas très net en renonçant à votre salle de sport. Je suis terriblement chanceux d'avoir pu partager tous ces moments, toutes ces discussions, toutes ces bêtises avec vous. Merci de m'avoir nourri intellectuellement et humainement, d'avoir encaissé mes obsessions et fluctuations émotionnelles, et pour votre réelle camaraderie. Merci aussi Camille pour tous tes tips de soin, les bons repas et la relaxation du visage!

Merci à tous mes amis d'avoir partagé les moments difficiles et de m'avoir aidé à sortir le nez du travail, à me détendre et à recharger les batteries.

Audrey, ton apparition soudaine dans ma vie a été une bouchée d'air frais qui m'a rappelé de respirer. Merci de m'avoir tenu en vie, de me montrer de nouveaux mondes et de me ramener à

d'autres que j'avais oublié.

Marine, merci pour toutes tes touches d'amour et d'attention qui m'ont permis de trouver la force et le sourire quand j'en avais le plus besoin (Bob a beaucoup aidé aussi!). Merci pour la main douce et patiente que tu m'as tendue et qui me fait tellement grandir.

Thank you, Mom and Dad, for being such wonderful and caring human beings. I am deeply grateful for your unconditional love and support, for fostering my curiosity about the world and for giving me this much freedom to figure out what to do with my life. I still don't know, but I try my best to embody the values you have taught me as I navigate this world – and it is quite the adventure! All my gratitude to my extended family as well for their continuous support (and sorry for missing your weddings because of the thesis kuzis...).

Finally, I owe my thanks to Lewis Pollock for his courageous and inspiring thesis. I discovered your work while desperately trying to make sense of the literature and of some data, and it completely changed my take on my dissertation. I hope you've found some peace of mind and sanity outside of academia.

Abstract

The view that cognition is embodied and that sensory-motor systems play a central role in its various processes has had a major impact on our current understanding of how knowledge is organised and represented. Despite having gathered substantial support over the years, this account has nevertheless been extensively debated and criticised. Some of the disagreements notably stem from empirical findings that appear difficult to reconcile with some of its core premises. The current work focuses on a subset of this literature that is frequently cited in support of multimodal knowledge representations, namely on studies investigating the role of motor information in the recognition and processing of manipulable objects. Through a broad and critical review, we argue that the evidence on this topic remains largely inconclusive, partly due to issues with methodological validity. The goal of this thesis is to question one widespread practice in particular: the use of Likert-type scale ratings (subjective variables) for stimulus selection and control. We start with a general investigation of the reliability and validity of subjective variables, and lay out a number of important implications for the studies using them. Armed with new insights about what such ratings represent, we then turn to how object manipulability has been operationally defined in the literature and offer an in-depth analysis and discussion of its different assessments. Finally, we present a new set of manipulability ratings for French words that we use to directly compare the results obtained through different rating instructions and to discuss their respective validities further. Overall, our results reveal a highly flexible and often inappropriate use of subjective variables, which both complicates cross-study comparisons and directly affects the validity of experiments. They additionally provide critical insights into the properties of Likert-type ratings, allowing us to propose recommendations for more robust practices and avenues for further inquiry.

Résumé

L'idée que la cognition est incarnée et que les systèmes sensori-moteurs jouent un rôle central dans ses différents processus a eu un impact majeur sur notre compréhension actuelle de l'organisation et de la représentation des connaissances. Bien qu'elle ait gagné un soutien important au fil des années, cette approche a également suscité de nombreux débats et critiques. Une partie des désaccords provient notamment de résultats empiriques qui semblent difficiles à concilier avec certaines de ses prémices centrales. Cette thèse se focalise sur une partie de cette littérature qui est fréquemment citée en faveur d'un ancrage multimodal des connaissances, à savoir les études sur le rôle des informations motrices dans la reconnaissance et le traitement des objets manipulables. A travers une revue étendue et détaillée, nous soutenons que les résultats sur ce sujet restent largement non concluants, en partie en raison de problèmes de validité méthodologique. L'objectif de cette thèse est de remettre en question une pratique répandue en particulier : l'utilisation de normes de type Likert (variables subjectives) pour la sélection et le contrôle des stimuli. Nous commençons par une investigation générale de la fiabilité et de la validité des variables subjectives, et présentons un certain nombre d'implications importantes pour les études qui les utilisent. À la lumière de ce premier travail, nous nous penchons ensuite sur l'opérationnalisation de la manipulabilité des objets et proposons une analyse et une discussion approfondies de ses diverses définitions dans la littérature. Nous présentons enfin de nouvelles normes de manipulabilité pour des mots français qui nous permettent de comparer directement les résultats obtenus à travers différentes consignes, et ainsi de complémenter nos analyses sur leurs validités respectives. Nos résultats révèlent globalement une utilisation très flexible et souvent inappropriée des variables subjectives, ce qui complique les comparaisons entre les études et affecte directement la validité des expériences. Ils fournissent en outre un éclairage essentiel sur les propriétés des évaluations de type Likert, nous permettant ainsi de proposer des recommandations pour des pratiques plus robustes et des pistes pour de futures recherches.

Contents

Acknowledgements **iii**

Abstract **vii**

Résumé **ix**

List of Figures **xiii**

List of Tables **xv**

Introduction **1**

1 Concepts, words, and objects **5**
1.1 Knowledge and its representation 6
1.2 Words and their meanings 16
1.3 A braid of action and perception 22

2 Evidence for the role of motor representations in object processing **33**
2.1 Affordances and compatibility effects 34
2.2 Object processing 41
2.3 Memory 53
2.4 The Body-Object Interaction effect 60
2.5 Discussion and thesis outline 64

3 The elephant in the middle of subjective rating scales **67**
3.1 The midscale disagreement problem 68
3.2 Low vs. High BOI 72
3.3 Regression studies 76
3.4 The present study 78

3.5 Methods . . . 78
3.6 Results . . . 80
3.7 Discussion . . . 86
3.8 Conclusion . . . 91

4 Through the forest of manipulability ratings **93**
4.1 A brief history of manipulability ratings . . . 95
4.2 The midscale disagreement problem in manipulability ratings . . . 98
4.3 Dimensions of manipulability . . . 100
4.4 Discussion . . . 120
4.5 Conclusion . . . 124

5 Dimensions of manipulability **125**
5.1 The present ratings . . . 126
5.2 Method . . . 128
5.3 Results . . . 131
5.4 Discussion . . . 143
5.5 Conclusion . . . 146

General Discussion **147**
6.1 Open questions and limitations . . . 150
6.2 Recommendations for unipolar Likert-type scale ratings . . . 155
6.3 A measurement crisis? . . . 160
6.4 Conclusion . . . 160

Bibliography **162**

Appendices **207**
Appendix A: Supplementary materials for Chapter 3 . . . 208
Appendix B: Supplementary materials for Chapter 4 . . . 210
Appendix C: Supplementary materials for Chapter 5 . . . 254

List of Figures

1 *The manual interference task's effect on naming errors against the subjective experience ratings in Yee et al. (2013)* 48
2 *The manual interference task's effect on errors (A) in Yee et al. (2013) and on latencies (B) in Davis et al. (2020) against participants' subjective manual experience ratings* 53
3 *Standard deviations of the BOI ratings as a function of their means for 9351 words collected by Pexman et al. (2019)* 69
4 *Living/Non-living (A; VanArsdall & Blunt, 2022) and age of acquisition (B; Kuperman et al., 2012) ratings against BOI ratings (Pexman et al., 2019)* 70
5 *Combined BOI ratings from Bennett et al. (2011) and Tillotson et al. (2008) against those provided by Pexman et al. (2019) for 1897 words in common* 71
6 *Classification accuracies in Pexman et al.'s (2017) concrete/abstract megastudy against the concreteness ratings provided by Brysbaert et al. (2014b)* 72
7 *BOI ratings of the stimuli used by a representative set of factorial studies against all the items in their reference datasets (left column: combined ratings from Bennett et al., 2011, and Tillotson et al., 2008; right column: Pexman et al., 2019)* 74
8 *BOI ratings of the stimuli used by a representative set of factorial studies as found in a different dataset than that which was originally used (left column: Pexman et al., 2019; right column: combined ratings from Bennett et al., 2011, and Tillotson et al., 2008)* 75
9 *BOI ratings (Pexman et al., 2019) of the words included in the analyses of Bennett et al., (2011), Hargreaves & Pexman (2014; Yap et al., 2012), Newcombe et al. (2012) and Taikh et al. (2015).* 77
10 *Histogram of the present BOI ratings (N = 1019)* 83
11 *Standard deviation as a function of the average BOI ratings in the present study (centre) and examples of item-level rating distributions* 84

12 *Standard deviation as a function of the average BOI ratings, along with item-level interrater agreement scores and type of rating distribution.* 86
13 *Standard deviation as a function of the average BOI ratings for words with an agreement score above .65 and absolute differences between the trimmed means and the average ratings* 87
14 *Standard deviations against the mean ratings for all identified manipulability-related datasets (left) and ridge plots for the distribution of the standard deviations for midscale items (right)* 101
15 *Standard deviations as a function of the average ratings and frequency distributions for the present norms* 139
16 *Correlation plot for the current and for our body-object interaction ratings* 141
17 *Standard deviations against the means for our combined ratings, divided by low ($<$ .65, left) and high ($\geq$ .65, right) agreement items* 152
18 *Averages and standard deviations derived from random sampling (N = 10,000) across varying sample sizes and source distributions* 154
19 *Global means against the trimmed means for items with an agreement score above .65 across all the unipolar variables collected in the current work* 159

List of Tables

1 Characteristics of the low- and high-BOI lists reported by the factorial design experiments 73
2 Results of the internal reliability analyses 82
3 Spearman correlation coefficients between the current ratings and those from other available datasets 82
4 Descriptive statistics for the present BOI ratings (N = 1019) and for the metrics used to assess their reliability. The absolute difference refers to the absolute value of the difference between each word's BOI rating and its trimmed mean 83

5 Excerpts from the instructions used to assess structural manipulability 103
6 Excerpts from the instructions used for – or related to – the assessment of functional manipulability 107
7 Excerpts from the instructions combining structural and functional manipulations . 119

8 Summary of the total and item-wise number of observations for each questionnaire 133
9 Results of the internal reliability analyses 133
10 Pearson correlation and linear regression coefficients between the current ratings and all other available datasets providing manipulability ratings 136
11 Descriptive statistics for the present ratings and for their agreement rates (N = 1019) 137

Introduction

Imagine being a researcher interested in how we come to know things and how this knowledge shapes our every-day interactions with the environment. For instance, when you see a cup filled with some black liquid on your desk, how do you know what it is, what to do with it, and how to do it? You have heard about an exciting hypothesis proposing that knowledge emerges from a reinstatement of past experiences, somewhat like sensory and motor 'simulations' based on previous interactions with objects. You decide to test this idea, and let us assume that you set out to investigate it through every-day manipulable objects, i.e. objects that can be grasped and used with the hands (e.g. the cup). Under your working hypothesis, knowing about such objects should involve, in part, an activation of the actions typically performed with them.

Having established the study's context, it is now time to operationalise the hypothesis. First, you would need to find a task that taps into the knowledge that people have of the objects. There is also the question of what the objects used in the experiment will be. For instance, if you chose to compare manipulable to non-manipulable objects, how would you select them? Would you simply list those that you can think of? Or perhaps you would like to make sure that participants really have experiences manually interacting with the manipulable objects, but not with the non-manipulable ones. One straightforward way to achieve this would be to ask a group of people before the experiment. Given that some objects might be more strongly manipulable than others, it also seems reasonable to collect graded responses on a Likert-type scale. But what would the question be? Would you ask them, for instance, to what extent they can physically interact with the objects (e.g. Siakaluk et al., 2008a)? Whether the objects are manipulable or not (e.g. Howard et al., 1995)? How much they are associated with manual actions (e.g. Carota et al., 2012)? Whether the hands are necessary for their function (e.g. Tranel et al., 1997)? After some consideration, you settle on one of these options and begin working on your rating task. A few additional questions arise at this point: among others, what should the list of items include, and how many participants would be enough? All these decisions and procedures are becoming a bit exhausting, and you would really just want to get started on your main experiment. So, you might include enough items that seem reasonably suited to your study's design – perhaps a few more, just to be safe or for

future use. You might also not need that many participants. After all, manipulable objects should be quite obvious to everyone. A quick tour of the lab or campus, and *voilà*! But it is not over. Now that you have gathered the data, you need to select appropriate stimuli for your experiment. Everyone computes averages, so you follow suit and derive a mean rating for each object. And now comes the tough part: deciding which ratings represent manipulable objects and which non-manipulable ones. Would you select them from the ends of the scale? What if the distribution of your ratings does not allow it? A median split, perhaps? You ponder the question for a while, and it seems reasonable that as long as non-manipulable objects have lower ratings than the manipulable ones, things should be fine. You create your two experimental lists and do a t-test to verify that they are significantly different: $p < .05$, done! All that now remains is deciding the details of the experimental task: number of trials, stimulus presentation time, stimulus size, response options...

However exaggerated, the above example illustrates very real decisions that must be faced in the course of preparing an experiment, often with no concrete resources or guidelines to inform them. As a result, researchers rely on vague intuitions or perpetuate common – but not necessarily well-founded – practices within their research teams or fields. Additionally, a large number of critical details may not be reported in the publications, not necessarily due to a lack of transparency, but simply because it would be impossible to describe and justify every choice – and likely quite overwhelming to read. The fundamental issue is that this gap in methodology introduces excessive flexibility in experimental practices, which can severely affect the validity of experiments and ultimately result in unreliable and confusing findings in the literature.

The goal of this thesis is to tackle some of these difficulties by gaining new and useful insights into stimulus selection procedures, as well as to help bring some clarity to discrepant findings. Our investigations will be grounded in the specific context of the conceptual representation of manipulable objects – very much in the sense introduced above. This topic has indeed produced a number of controversial results that are theoretically difficult to interpret and have hampered progress, which makes it an excellent case study to investigate the potential impact of methodological problems. In Chapter 1, we provide the core theoretical and methodological frameworks of this literature. We then turn to a critical review of the empirical evidence on the role of motor information in the processing of manipulable objects, which allows us to highlight more specific issues, namely how we define and assess manipulability (Chapter 2). Regarding our investigations, Chapter 3 first offers a detailed investigation of the tool typically used to assess stimulus characteristics: the Likert-type scale. In Chapter 4, we offer a review of how manipulability has been operationally defined, and use our previous findings to evaluate the validity of its different assessments. Following our conclusions about the tasks that appear the most appropriate to capture manipulability, we then present a new set of ratings for French words and further examine their

validity (Chapter 5). The thesis finally concludes with a list of preliminary recommendations for the collection, interpretation and use of Likert-type ratings, and highlights several questions that remain to be addressed.

Chapter 1

Concepts, words, and objects

As discussed in the Introduction, the driving motivation behind this thesis is to investigate our stimulus selection procedures. Such an inquiry nevertheless requires to be conducted in context – not only to understand the rationale and methods of current practices, but also to better assess their impact on the validity of experiments. For our current purposes, this context will be the representation of knowledge, and particularly that of manipulable objects. The current chapter lays the theoretical and methodological foundations of this topic. In Section 1.1, we will first present the broad theoretical framework of conceptual representations. The two subsequent sections will focus specifically on the primary ways through which our knowledge of objects can be effectively investigated: their names (Section 1.2) and their visual presentation (Section 1.3). Each section provides a historical perspective on its respective literatures and ends with a discussion of current approaches and methodological considerations. Given the breadth of topics that will be covered, some perspectives will necessarily be omitted or only briefly mentioned despite their importance. This should nevertheless not be a major concern as our primary focus will be methodological rather than theoretical.

1.1 Knowledge and its representation

1.1.1 Descartes in the information age[1]

Modern cognitive science emerged in the midst of major intellectual and technological developments that profoundly shaped how the mind and the brain were conceived. Spearheaded by information theory (Shannon, 1948), the advent of computing machines (Turing, 1937; von Neuman, 1993) and the first wave of artificial intelligence (e.g. McCarthy et al., 2006), the new paradigm firmly took root in the idea that brains are information processing machines. To illustrate this view, Wheeler (2005) gives the example of the *sense–model–plan–act* (SMPA) framework taken from robotics (Brooks, 1991), which he describes as "the intellectual core of orthodox cognitive science" (p. 68). In the SMPA approach, the brain first receives information from the environment through its receptors. The signal is then transduced into a symbolic format – much like a language – that can be "read" by the cognitive system[2] (Fodor, 1975, 1983; Pylyshyn, 1973. See also Newell, 1980), providing an objective description (or representation) of the world (see Anderson, 2017). Once this model has been formed, it can be computationally operated on using instructions and knowledge

[1] For a presentation of the rationalist and cartesian foundations of traditional cognitive science see, e.g. Wheeler (2005) and Dreyfus (1988, 1995).

[2] As stated by Fodor (1983), *"what perception must do is to so represent the world as to make it accessible to thought"* (p. 40, emphasis in original)

from memory in order to reason and plan appropriate actions. The process finally results in a motor command and overt behaviour. This framework thus essentially describes the brain as a serial computer: sensation input → representation → computation → output. In the words of Jerry Fodor and Zenon Pylyshyn (1988), two of the most prominent proponents of this approach, "[c]lassical models of the mind were derived from the structure of Turing and Von Neumann machines. They are not, of course, committed to the details of these machines as exemplified in Turing's original formulation or in typical commercial computers; only in the basic idea that the kind of computing that is relevant to understanding cognition involves operations on symbols" (p. 4).

This approach inevitably led to the conception of the brain as made up of highly specialised and mostly independent systems (or modules) – somewhat similar to the components of a computer or to the organs of the body (Chomsky, 1980; Fodor, 1983. For examples of such architectures see, e.g. Anderson, 1996; Anderson et al., 2004; Atkinson & Shiffrin, 1968; Newell, 1980). Among these, semantic memory[3] is seen as the storage (or, as Tulving put it, "mental thesaurus", p. 386, 1972) of conceptual knowledge. Most early work described concepts as being defined by discrete properties and aimed to identify the type of symbolic structure that could support cognition. For instance, Rips et al. (1973; Smith et al., 1974) proposed that concepts are represented as lists of features that are sorted by how defining and characteristic they are. To take an example from the authors, the concept of *robin* would include features such as being bipedal, having wings, having distinctive colours, perching trees and being undomesticated. Quillian (1967. See also Collins & Quillian, 1975) instead envisioned a semantic network model in which each concept is represented by a node and related to its properties (i.e. other concept nodes) by labelled links. The node *robin* would thus be associated with, e.g., the node *bipedal* by the link *is*. Other authors introduced yet different organisations, such as frames (Minsky, 1974) and scripts (Schanck & Abelson, 1975). All of these models have in common that they treat knowledge as represented in an abstract format and within language-like syntactic structures. As Barsalou (1999) points out, these symbols are also "amodal because their internal structures bear no correspondence to the perceptual states that produced them" (p. 578), thus implying that they are arbitrarily linked to them. Much like the word *hammer* has no similarity to our perceptual experiences and interactions with the physical object, its concept in memory is also effectively detached from what it represents.

[3]Early work in cognitive science mainly conceived memory as a unitary sytem. The results of a large number of behavioural, neuropsychological and animal studies over the second half of the 20th century nevertheless led to important dissociations and to postulate multiple memory systems (see Baddeley, 2018; Eichenbaum, 2010; Squire, 2004. For a critical view, see Versace et al., 2009).

How does meaning arise in such a symbolic system?

Imagine that you are learning a new language, say Greek. If you translated Greek words back to your first language, you could understand what they mean (e.g. that σφυρί means *hammer*) – i.e. because you understand your first language. Looking up their definitions in a Greek-Greek dictionary or translating them into a language that you do not know would not get you very far. But then how do you understand your first language? The symbolic account essentially amounts to saying that this happens through a translation from first language to another, 'inner' one. This leaves us facing the same problem we had started with: we now need to explain how we understand the inner language (Dummett, 1993). As Margolis and Laurence (2023) put it, "the mental representation itself is just another item whose significance bears explaining. [...] [W]e are involved in a vicious regress, having to invoke yet another layer of representation (and so on indefinitely)". Proposing language-like symbolic structures to account for conceptual knowledge thus does not bring any added explanatory value; it just displaces the problem without resolving it. This little thought experiment illustrates what is generally known as the *symbol grounding problem* (Harnad, 1990. See also Searle, 1980). It challenges the idea that a symbol system can be intrinsically meaningful when its symbols are themselves defined only in terms of other symbols, without ultimately being grounded in what they represent.

1.1.2 Distributed representations

The dominant approach to the organisation of conceptual knowledge presented above was mainly motivated on theoretical grounds and investigated through the ability of computational models to predict the results of behavioural experiments. The question of how conceptual knowledge is represented at the neural level was nevertheless largely left unaddressed. Two major lines of work in the 1980s bridged this gap and laid the foundation of our current understanding of how knowledge is organised: cognitive neuropsychology and connectionism.

Category-specific semantic deficits

The idea that the brain is composed of modular subsystems sparked a 'hunt' for dissociations in patients with acquired neurological disorders. Dissociations occur when a patient with a brain lesion shows impaired performance on one type of task (A) but not another (B). Given the modularity assumption, it can be postulated that the lesion has affected the subsystem whose function underlies the realisation of task A. The hypothesis is greatly corroborated if a double-dissociation is found, with another patient displaying the reverse pattern of results (i.e. spared performance on task A and impaired on task B). Double-dissociations are generally taken to show that two separate

subsystems are at play and that the results are not due to methodological factors (We will see other examples of dissociations in Section 1.3. For a general discussion, see also, Shallice, 1988).

In a series of seminal studies, Elizabeth Warrington, Tim Shallice and Rosaleen McCarthy established that brain damage can result in dissociations between different semantic categories. Warrington and Shallice (1984) reported four patients who had large and disproportionate impairments in the recognition and description of animals, plants and foods compared to inanimate objects[4]. Although the experimental tasks somewhat differed, the opposite pattern of results (relatively impaired recognition of inanimate objects compared to animals, plants and foods) was also observed in another patient described by Warrington and McCarthy (1983). Instead of postulating the existence of category-wise subsystems, the key insight of the authors was that the results might reflect the type of information involved in the representation of the categories (for the former view, see Caramazza & Shelton, 1998. For reviews on different theoretical stances, see Cree & McRae, 2003; Mahon & Caramazza, 2009). This view holds that the apparent deficits are not related to specific categories as such, but rather arise from the sensory and functional attributes of objects that are most relevant to their recognition. Animals and plants, for instance, are mostly characterised by their visual features, whereas inanimate objects are generally related to specific functions and uses. Assuming modality-specific semantic subsystems, a lesion preventing access to, e.g., stored visual attributes would greatly impair the recognition of animals and plants, but to a much lesser extent the recognition of inanimate objects. This sensory/functional theory thus postulates that conceptual knowledge is distributed across modality-specific semantic subsystems and thus that the contents of conceptual representations are sensory and functional attributes. Note, however, that what 'functional attributes' refer to is not entirely clear under this account.

In a subsequent case study, Warrington and McCarthy (1987) provided evidence for – and predicted – finer-grained categorical dissociations based on their model. One of this study's findings is particularly relevant for the present purposes. Within the broader domain of manmade objects, the patient's performance was impaired only for relatively small and manipulable objects (e.g. kitchen utensils, office supplies, furniture). This finding lead to a further refinement of functional features into motor attributes related to object use and knowledge of function – which was itself postulated to rely on sensory-motor information. It is apparent from these authors' work that their position is at the fringes of the modular and symbolic approach. Warrington and McCarthy (1987)

[4]Two of their patients were tested on naming and describing visually presented objects, as well as giving the definitions of auditorily presented words. Two others had important language impairments that did not allow them to perform these tasks. They were tested through a matching paradigm in which they presented with five pictures and needed to associate a spoken word with the correct one. Note also that the authors performed a number of different experiments and that not all participants were tested on the mentioned categories (particularly plants).

particularly noted that their "current perspectives have much in common with contemporary computational connectionist models of knowledge processing in which representations are computed on the basis of weighted entries in parallel distributed processing systems" (p. 1292).

Connectionism

Connectionism provides a fundamentally different approach to the study of cognition than the classical modular view. Although its inception dates back to the 1940s, the movement only gained substantial prominence in the early 1980s[5] (e.g. Rumelhart et al., 1986. For a historical take, see Medler, 1998; Pollack, 1989). The general framework can be roughly summarised as the attempt to understand cognitive processes as emerging from the parallel and distributed activity of neural networks[6]. These are typically modelled as interconnected units (the neurons), where each connection (akin to a synapse) has a weight indicating its strength and whether it is inhibitory or excitatory. As the network receives inputs, the weights get gradually adjusted so that concurrently active units strengthen an excitatory connection, while inactive units get increasingly inhibited (similar to Hebbian learning. Hebb, 1949). In essence, such a network thus learns statistical regularities (patterns) in its inputs. The critical difference with the modular approach is that what the network 'knows' is not something that is passively stored in a specific location, waiting to be retrieved. Rather, it is implicit in – and distributed across – the entire network's connectivity pattern, and thus integral to how information is processed (for a non-technical presentation of some interesting properties of such networks, see Allport, 1985; Clark, 2001).

1.1.3 The embodied mind

The fundamental issue with the classical computational approach is that it builds on the assumption that thinking, reasoning, problem-solving, planning, etc. – in short, 'higher cognition' – are the characteristic features of human-level intelligence and cognition. Trying to model and to understand these processes is not a problem in itself, of course. The danger, however, is in mistaking them for the primary function of the brain. A major criticism of this approach has been that it gives a very distorted view of what cognition is and of what brains are for, marginalising essential aspects of what being a living thing in the world entails. Besides, robotic and artificial

[5] One of these is likely the failure of symbolic artificial intelligence (AI) systems to meet their promises, and particularly the difficulties encountered while trying to implement symbolic architectures. I highly recommend the documentary film *Being in the world* (Ruspoli, 2010) which presents these issues and offers a phenomenological critique of the symbolic-computational approach.

[6] It should be noted, however, that connectionism does not necessarily exclude the possibility of symbol systems. Indeed, a connectionist network can in principle be used as an implementation of a functionally symbolic structure (e.g. Smolensky, 1988).

intelligence systems built on the classical model of the mind simply did not work except in specific and narrow cases. They were too rigid and static, incapable of the dynamic and flexible behaviours that we and other animals display (see e.g., Anderson, 2003; Brooks, 1991; Dreyfus, 1992).

Beginning roughly in the 1990s, a movement across virtually all fields of cognitive science started drawing attention to our embodied nature and embeddedness in the world[7] (e.g. Barsalou, 1999; Brooks et al., 1991; Clark, 1998; Glenberg, 1997; Lakoff & Johnson, 1980; O'Regan & Noë, 2001; Varela et al., 1993. For this view's historical roots, see Barsalou, 2010; Glenberg et al., 2013; Shapiro & Spaulding, 2021). At its core, this view rejects the portrayal of cognition as detached reasoning about an independent external world. Rather, it is argued that cognition must be understood as fundamentally shaped by – and grounded in – the context of a biological body that navigates and interacts with a rich and dynamic environment. Although different theoretical flavours have sprung from this *embodied cognition* movement (see Jacobs, 2015; Shapiro, 2011; Wilson, 2002), their general rallying point is that perception and action are integral to cognition, as opposed to peripheral input and output systems. There has nevertheless been a rift within embodied cognition regarding the adherence to the very notion of mental representations – sometimes called the 'representation wars' (Clark, 2015). The enactivist thesis generally advocates for a dynamic systems approach and rejects a recourse to representations to explain cognitive phenomena (e.g., Buhrmann et al., 2013; Chemero, 2013; Hutto & Myin, 2014; O'Regan & Noë, 2001. See also footnote 11). Less radical views have continued to rely on representations, but offer a quite different perspective on their nature compared to the classical approach. For our present purposes, we shall remain unabashedly representational.

Representation as multimodal simulation

Allport (1985) was arguably one of the first to offer what could be called an embodied account of conceptual knowledge. Based on a connectionist architecture, he proposed a distributed model of semantic memory in which the same sensory-motor networks that mediate our interactions with objects are also involved in their representation. However, it is not until Barsalou's (1999, 2008; Barsalou et al., 2003) influential work that this idea was elaborated into a fully-fledged conceptual system (often referred to as *grounded cognition*). The gist of the model can be illustrated as follows. As we engage with the world, our neural systems gradually 'tune in' to the regularities in our experiences. They learn recurrent patterns of activity both within and between sensory-motor modalities, as well as how they relate to the organism (e.g. shapes, sounds, smells, colours, their associations and dynamics, how they affect internal states, how they change as we

[7] There were, of course, important precursors to the movement. We will briefly present one of them, James J. Gibson's ecological approach, in Section 1.3.2

move different parts of our bodies, etc.). By tracking and integrating these regularities, the brain's structure essentially mirrors the organism-relative properties of the world, in a sense becoming an embodied 'model of the world'. In turn, such an architecture can now enact the world, i.e. simulate experiences that could, in principle, result from real interactions. Barsalou (1999) more specifically argued that multimodal neuron assemblies in the brain's convergence zones (Damasio, 1989) encode a subset of active perceptual, motor, and introspective states during our experiences. Common activity patterns form 'simulators' in these areas, enabling flexible sensory-motor re-enactments based on previous experiences.

As in our discussion of connectionism (Section 1.1.2), knowledge is here seen as embedded in and distributed across the brain. More importantly, its 'retrieval' requires multimodal simulations (often unconscious) akin to a partial re-enactment of previously experienced states. Note that this view offers a radically different account of representations than the classical approach. In contrast to amodal symbols on which cognition operates, simulations are directly grounded in the perceptual, motor and introspective states that arise from our experiences. The key hypothesis is thus that the fundamental 'currency' (or format) of representations and of cognitive activity more generally is sensory-motor information. The model also forces a re-examination of what concepts are. Indeed, it does not posit the existence of discrete and well-defined concepts. Rather, it describes a mechanism through which we produce "context-specific representations of a category" (p. 84, Barsalou et al., 2003) – never in isolation and virtually always involving knowledge about associated entities and events.

Similar accounts to the grounded view have emerged across a variety of fields, postulating shared processing resources between cognition and sensory-motor systems (e.g. Decety, 1996; Gallese, 2007; Gibbs, 2006; Glenberg, 1997; Grush, 2004; Hesslow, 2002; Jeannerod, 2001; Pulvermüller, 1999; Schacter et al., 2007). A large body of literature generally supporting this hypothesis has also accumulated over the past three decades (for reviews, see e.g., Barsalou, 2008; Brunel et al., 2015; Glenberg, 2015; Kiefer & Pulvermüller, 2012; Martin, 2007, 2016; Muraki et al., 2023a; Pearson & Kosslyn, 2015; Pecher, 2013a; Pexman, 2019). To pick just a few examples, several studies have shown a close correspondence between both behavioural and neural activity during the 'encoding' and the 'retrieval' of information (Danker & Anderson, 2010; Kent & Lamberts, 2008). Neuroimaging studies have also revealed the involvement of sensory-motor areas in a variety of tasks. For instance, reading words strongly associated to odours activates the olfactory system (González et al., 2006), and areas involved in colour perception overlap with those active during judgements about the colour of a word's referent (Hsu et al., 2012; Simmons et al., 2007). On the 'motor' side, reading action words or sentences involving different parts of the body was found to result in a somatotopic activation of the motor cortex (Hauk et al., 2004; Tettamanti et al.,

2005). Conversely, immobilising the dominant hand for 24h perturbs judgements on hand-related action words (Bidet-Ildei et al., 2017) and alters distance perception in peripersonal space (Toussaint et al., 2020). One particularly telling set of results suggests that both imagining and viewing images of foods activates the gustatory cortex, with a representation of specific taste qualities (e.g. sweet, salty. Avery et al., 2020, 2021, 2023). Most interestingly, the activity in the gustatory cortex while viewing pictures of foods was found to be modulated by the glucose level in participants' blood: the lower the glucose level (and thus the energy resources available to the body), the higher the activity in the gustatory cortex (Simmons et al., 2013). Overall, these and a vast amount of other results point to a sensory-motor representational format and to a highly context-sensitive conceptual system. However, the debate about how knowledge is organised and represented is far from settled.

1.1.4 Debates and challenges

Despite its wide adoption and experimental success, the modality-specificity of conceptual representations has also attracted a fair share of criticism. Proponents of an amodal representational format have pointed out that a large portion of the evidence cited in support of the grounded view can just as well be explained by spreading activations from an amodal conceptual system. In their *grounding by interaction* approach, Mahon and Caramazza (2008. See also Mahon, 2015; Mahon & Hickok, 2016) have argued that activations from amodal conceptual representations can spread to sensory-motor areas, enriching the representations and complementing processing. However, this does not imply that sensory-motor information is constitutive of a given concept, nor that it necessarily plays a functional role in its processing. Another issue presented by these authors is more particularly directed towards neuroimaging studies. They have argued that an overlap of brain activity between perceptuo-motor and conceptual processes can be due to the activation of abstract representations stored in regions adjacent to sensory-motor areas (the *anterior shift hypothesis*). It is thus possible that neuroimaging data have been falsely interpreted as reflecting modality-specific activations. On a related point, the general assumption that some cortical regions process modality-specific information (e.g. visual, auditory, motor) has itself been questioned. Calzavarini's (2023) recent review suggests that these regions might instead be processing specific properties (e.g. motion, facial information, shape) independently of the nature of the incoming sensory information. If such an account turns out to be the case, it would require a serious re-examination of past neuroimaging findings. The general difficulty posed by these questions for any theory was perhaps best captured by Martin (2016) who notes:

> The problem, however, is that we do not know how to determine the format of a representation (if we did, we would not still be debating the issue). [...] What we do know is that at the biological level

> of description, mental representations are in the format of the neural code. No one knows what that is, and no one knows how it maps onto the cognitive descriptions of representational formats (i.e., amodal, propositional, depictive, iconic, and the like), nor even if those descriptions are appropriate for such mapping. (p. 984)

The functional role of multimodal simulations in conceptual processing has also been challenged by a number of inconsistent results. Ostarek and Bottini (2021) recently argued that the field should move towards stronger inferential paradigms and provided a review of such evidence from congenital sensory-motor disorders, acquired sensory-motor deficits and behavioural interference paradigms. In all cases, results appear to be largely inconclusive, partly due to methodological confounds, differences in experimental protocols, and sometimes a failure of the theory to provide explicit predictions (see also Kaschak & Madden, 2021). For instance, Ostarek and Huettig (2019) have noted that opposite experimental findings are sometimes equally taken as evidence for multimodal simulations, which naturally poses an epistemological problem: if any result can be explained with the same claim, then none can falsify it (for a related point, see Mahon & Hickok, 2016). Some studies have used transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS) techniques to non-invasively interfere with specific perceptual or motor brain regions during conceptual processing. Unfortunately, these have also yielded contradictory results (for reviews, see Ostarek & Bottini, 2021; Papeo et al., 2013). Critically, there is evidence that this subfield suffers from low statistical power and strong publication bias (Solana & Santiago, 2022, 2023), while it remains to be determined if the rest of the literature on multimodal simulations stands on solid ground. Recent findings – coupled with what we already know of the psychological and neuroscientific literatures (e.g. Button et al., 2013; Ioannidis, 2005; Open Science Collaboration, 2015; Scheel et al., 2021) – are nevertheless not very encouraging. Replication attempts of some key findings in the field have yielded null results (e.g. Chen et al., 2024; Colling et al., 2020; Montero-Melis et al., 2022; Morey et al., 2022; Petrova et al., 2018; Saccone et al., 2021), and the interpretation of others as reflecting multimodal simulations did not hold under more robust analyses or paradigms (e.g. Ostarek et al., 2019; Witt et al., 2020).

A possible explanation for the seemingly contradictory findings is that conceptual processing might not activate all knowledge about a concept automatically, but rather flexibly draw upon relevant information depending on the situation at hand (e.g. Casasanto & Lupyan, 2015; Lebois et al., 2015; Yee & Thompson-Schill, 2016. For a cautionary take, see Mahon & Hickok, 2016). Indeed, the implicit assumption of a large number of experiments has been that conceptual processing leads to a simulation of all sensory-motor aspects associated with a given concept. If simulations are context-dependent, however, this could explain why some modality-specific effects are not consistently observed across experiments. Additionally, the role of linguistic information in conceptual

processing has been largely ignored in early research. In line with a flexible conceptual system, several authors have argued that the statistical regularities in language could contribute significantly to various tasks, reducing the need for multimodal simulations of entities and events (e.g. Barsalou et al., 2008; Cayol & Nazir, 2020; Connell & Lynott, 2014; Johns & Jones, 2012; Kemerrer, 2015; Dove, 2011, 2022; Louwerse, 2008, 2018. See also Section 1.2.2). Although highly pertinent and insightful, this view nevertheless also comes with some baggage. It implies a meticulous selection and design of experimental tasks and, more importantly, precise predictions based on the specific processes (and their dynamics) that the tasks tap into. In other words, it requires the validity and a mastery of the tools that we use to probe conceptual representations (Zwaan, 2021). An additional layer of complexity comes from the fact that different participants might rely on different processes to perform the same task. There thus appears to be a need to develop tools that reliably assess individual differences and strategies, which could also help elucidate the findings in congenital or acquired deficits and disorders (see e.g., Barsalou, 2016; Kemerrer, 2015; Ostarek & Bottini, 2021).

1.1.5 Synthesis

The embodied account of cognition has emerged as a highly influential framework and describes conceptual representations as flexible simulations that draw on sensory-motor processes. For our current purposes, the general hypothesis regarding manipulable objects that stems from this view is that they are represented, in part, by the actions typically performed with them. Beyond overtly action-oriented contexts, however, under what conditions these actions would play a role in their conceptual processing is less clear.

While this approach has gathered considerable support, we have seen that it also remains highly debated, partly on epistemological grounds and partly because of inconsistent empirical findings. We will further explore these points in the context of manipulable objects in Chapter 2 and show that the evidence is largely inconclusive on whether their representation involves a simulation of their associated actions. Before we can delve deeper into this topic, we nevertheless need a better understanding of the methodological and theoretical frameworks that have shaped the study of conceptual knowledge and manipulable objects. In the following section, we turn to psycholinguistics and to the study of word meanings. In Section 1.3, we will more specifically focus on the processing of manipulable objects from a neurocognitive perspective.

1.2 Words and their meanings

Early studies and models of visual word recognition primarily aimed to identify how the properties of words – from basic visual features to incrementally higher levels (e.g. letters, phonemes, graphemes, morphemes, meaning) – contribute to their recognition (for reviews, see Balota, 1994, 2006; Rastle, 2016; Yap & Balota, 2015; Yelland, 1994). This literature has notably yielded an abundance of findings regarding the influence of variables at the level of whole words (e.g. their number of letters, their frequency of occurrence in language), which will be our main focus in this section.

1.2.1 Visual word recognition

An important characteristic of this line of research is that there has been a plethora of experiments but using a rather small number of simple tasks, which greatly facilitates the comparison of the findings across studies. Two widely used procedures, for instance, are the naming task (i.e. word pronunciation) and the lexical decision task (LDT) in which participants must decide by pressing one of two keys if a string of letters corresponds to a real word or not (e.g. *book* and *pokf*). The time participants take to respond to each word (and their accuracies) is considered a reflection of the processes underlying word recognition, and explaining the differences in performance between words has largely been the 'name of the game' in this literature. Numerous experiments have established a number of important variables affecting how easily a word is recognised. One of the strongest predictors has been consistently found to be a word's frequency as derived from its number of occurrences in text corpora, with faster responses observed for high-frequency words (e.g. Brysbaert & New, 2009; Stanners et al., 1975; Whaley, 1978). Among others, some extensively studied variables also include the length of words (New et al., 2006), their orthographic similarity to other words (Coltheart et al., 1977; Yarkoni et al., 2008) and the age at which they are acquired (Gilhooly & Watson, 1981; Johnston & Barry, 2006; Juhasz, 2005). The findings relative to these variables – and to word frequency in particular – have been central to most early models that tried to capture how visually presented words are mapped onto a lexical representation (*lexical access*), thus allowing their identification and the production of a response (e.g. McClelland & Rumelhart, 1981; Morton, 1969).

The quest to identify the variables predictive of word recognition performances across a limited number of tasks has had several interesting consequences. For instance, it became apparent quite early on that the mentioned variables yield different effects across tasks presumed to tap the same cognitive processes. In a seminal study showing such differences, Balota and Chumbley

(1984) notably proposed that lexical processing is much more flexible than described by the dominant word recognition models of the time (see also Balota et al., 1991). Imagine for instance that you are performing a LDT. Some real words that are very familiar and meaningful to you will be quite easy to be identified as such (e.g. *phone*). Some letter strings will also be too strange to be real words (e.g. *agink*). In some cases, however, a pseudoword might seem quite similar to a real one (e.g. *chumingly*), whereas an unfamiliar and not very meaningful real word (e.g. *ortolidian*) will likely be more difficult to recognise. A quick verification in such cases is not enough and requires deeper processing to decide. Balota and collaborators thus argued that word recognition tasks – and the LDT in particular – do not only require lexical access as it was generally assumed, but that they can also involve access to a word's meaning.

1.2.2 Semantic richness & types of semantics

The proposal that meaning can influence a word's recognition initially received little attention but had some empirical support. A prime example of this is *concreteness*, a well-established variable in verbal memory studies following the influential work of Allan Paivio (for reviews, see Paivio, 1971, 1999). Concreteness captures the extent to which a word's referent can be experienced through the senses[8] (Spreen & Schulz, 1966; Paivio et al., 1968) and was known to have a facilitatory effect in verbal memory. A few early studies had suggested that a word's concreteness can also play a role in word recognition (e.g. James, 1975; Rubenstein et al., 1970) but the claim remained largely controversial due to methodological limitations and to potential confounding factors (e.g. Gernsbacher, 1984). In a later review on the role of such semantic dimensions, Balota et al. (1991) convincingly argued that the extant literature points to the conclusion that "*more-means-better*" (p. 214, emphasis in original) – i.e. that "words with more meaning representations are recognised more quickly" (p. 215).

With the attention turning to semantics, a new question quickly emerged. As Connell & Lynott (2015) noted, "[w]ords have meanings; on that much, psycholinguists generally agree. However, the issue of what 'meaning' is, and why a word's semantic content affects how easily it is recognised, are matters of less consensus" (p. 71). How should meaning be operationalised? The first studied variables were generally borrowed from research areas focusing more heavily on semantics (and influenced by the models of conceptual knowledge of the time, see Section 1.1.1)

[8]Concreteness is a subjective (or semantic) variable obtained by asking a group of participants to rate words on a Likert-type scale. Low values correspond to abstract concepts (e.g. *justice*) that are not associated with any sensory experience, while high values capture concrete entities (e.g. *hammer*). This variable has often been used interchangeably with a highly similar and correlated subjective variable: imageability. The latter captures how easily a given word can evoke a mental image (for discussions about the difference between the two variables, see Bonin et al., 2018; Connell & Lynott, 2012; Dellantonio et al., 2014).

and were constrained to some extent by the available normed datasets. For instance, Balota et al. (2004) investigated variables such as imageability, the number of different associates produced to words in a free association task (see Nelson et al., 2004) and the number of connections words have with other words in large networks such as WordNet (Miller, 1990; Steyvers & Tenenbaum, 2005). All variables predicted both naming and LDT latencies after controlling for other critical dimensions. Two other notable semantic variables found to facilitate word recognition were the number of features that are listed in response to a given word (e.g. Pexman et al., 2002; Grondin et al., 2009) and its semantic neighbourhood derived from statistical co-occurrences in text corpora (e.g. Buchanan et al., 2001; Pexman et al., 2008. For a review of such facilitatory *semantic richness* effects, see Pexman, 2012).

At around this time, embodied accounts of cognition were also gaining considerable traction and brought new insights about what the 'meaning' of a word might entail. This sparked considerable interest in variables aimed at assessing the experiential aspects associated with words' referents. For instance, body-object interaction (BOI, Siakaluk et al., 2008a) ratings reflect the ease with which the human body can physically interact with an object and can be seen as a coarse measure of its association to motor information (we will discuss this variable at length in Section 2.4 and in Chapter 3). On the other hand, Juhasz et al. (2011, 2013, 2015) assessed how much sensory experience is associated with what words represent. Taking a finer-grained approach, Amsel et al. (2012. See also Medler et al., 2005) collected norms on several sensory-motor dimensions such as the extent to which entities can be grasped and are associated with motion, colour, smell, sound and pain. Connell and Lynott (2012; Connell et al., 2018; Lynott et al., 2020) similarly provided ratings for the strength of association to five senses, five motor effectors and to interoception. In most cases, these experiential dimensions have been shown to have facilitatory effects on word recognition performances after controlling for a number of important variables, thus bringing support to the sensory-motor grounding of conceptual knowledge.

The introduction of semantic variables and embodied perspectives has also fostered inquiries into tasks requiring deeper semantic processing, notably semantic classification tasks (SCT). These typically require participants to decide to which of two semantic categories words belong (e.g. is *cloud* concrete or abstract? living or non-living?). It has been shown that the effects of semantic variables are typically more pronounced in SCTs than in shallower tasks such as the LDT or naming (e.g. Heard et al., 2019; Pexman et al., 2019; Yap et al., 2012. See also Section 2.4). Additionally, these effects have been found to vary across different decision requirements in the SCTs, thus reinforcing the view that word processing flexibly draws on different types of information depending on the situation at hand. This line of research has mostly converged with the so-called hybrid accounts in embodied cognition that posit two complementary representational systems (see also

Section 1.1.4). One of them is the grounded experiential system discussed in Section 1.1.3 which can re-enact perceptual, motor and introspective states. The second is a linguistic system specifically tuned to the statistical regularities in language, which can act as a 'shortcut' (Lynott & Connell, 2010) or a 'quick-and-dirty' way (Louwerse, 2018) for deriving meaning. Tasks that do not require highly detailed conceptual processing could thus be solved by the latter system, while deeper processing would involve situated experiential simulations. The relative contributions of these two systems to conceptual processing – and how they dynamically interact – are nevertheless not fully understood and remain active areas of research (e.g. Brown et al., 2023; Petilli et al., 2021; van Hoef et al., 2023; Wingfield & Connell, 2022).

1.2.3 Big data psycholinguistics

An important detail which we have glossed over and that has made all these inquiries possible is the shift in the word recognition literature from factorial design studies to regression analyses, and particularly the introduction of *megastudies*. Word processing – as many other processes in cognitive psychology – has been typically studied by dichotomising variables, i.e. by creating separate lists of words which are equated on variables known to affect recognition performances, and differing along the variables of interest. As the number of variables shown to affect word recognition performances grew, it became exceedingly difficult to reliably control all relevant dimensions and to find adequate stimuli that fit a study's criteria. This situation was nicely captured by Anne Cutler's (1981) commentary, with its witty title: *Making up materials is a confound nuisance, or: Will we be able to run any psycholinguistic experiments at all in 1990?* Such factorial designs were additionally increasingly criticised due to a number of issues including – but not limited to – experimenter biases in stimulus selection, low statistical power and loss of information regarding the variable's effect on its continuum. Finally, a notable limitation is that factorial designs only allow to conclude whether an effect is statistically present or not, providing little information about the relative importance of different variables (for more comprehensive discussions see, e.g. Baayen, 2010; Balota et al., 2012; Brysbaert et al., 2014a, 2016).

With advancements in technology and thanks to the relatively simple experimental procedures used in the field (especially naming and the LDT discussed previously), researchers started conducting *megastudies*, providing large behavioural datasets with summary statistics for several thousands of words (e.g. Seidenberg & Waters, 1989; Balota et al., 2004. For a historical review, see Keuleers & Balota, 2015). As efforts to norm large numbers of stimuli had already started (e.g. Coltheart, 1981; Cortese & Fugett, 2004; Ferrand et al., 2008; Gilhooly & Logie, 1980; New et al., 2004; Togglia & Battig, 1978), these megastudies made it immediately possible to conduct

large-scale regression analyses to gain new insights into how (and to what extent) different variables affect word recognition performances. A significant advance in this approach was brought by Balota et al.'s (2007) English Lexicon Project with LDT and naming performances for approximately 40,000 words, shortly followed by similar undertakings in other languages and with a variety of experimental tasks (e.g. Ferrand et al., 2010, 2011, 2018; Goh et al., 2020; Keuleers et al., 2012; Pexman et al., 2017; Sze et al., 2014; Yap et al., 2010). In turn, this has stimulated the collection of even larger norming datasets with ratings for several thousands to tens of thousands of items (e.g. Brysbaert et al., 2014; Kuperman et al., 2012; Lynott et al., 2020; Pexman et al. 2019; Wang et al., 2023; Winter et al., 2024). The availability of megastudies has been instrumental for testing the effects of new variables, and especially those related to the experiential aspects of concepts discussed earlier.

1.2.4 Open questions and challenges

A large number of studies on word recognition and processing relies on the methodology presented in section 1.2.1, i.e. on whether various word-level variables predict differences in task performance (reaction times and accuracies). The inference of cognitive processes and representational format through this approach nevertheless requires several leaps of faith, especially for semantic dimensions. One notably has to assume (1) that a given variable does indeed capture the targeted property of concepts, and (2) that its effect implies that the same type of information captured by the variable is involved in the processes engaged during the task (and conversely, that a lack of an effect reflects that it plays no role). To take an example, let us suppose that we have quantified the manipulability of objects by asking participants to rate on a scale how much they associate manual action with different objects. We assume that the obtained values fall on a continuum from low to high manipulability – and nothing fundamentally different (1). We then run an experiment (or use the data from a megastudy) and perform analyses to determine if changes in manipulability explain changes in performance. If, for instance, a facilitatory effect of the variable is observed (i.e. better performance for highly manipulable objects compared to low), the typical conclusion would be that manual motor information contributes to the task, and that this reflects the sensory-motor grounding of semantic knowledge (2).

The first hurdle for postulating (2) is that the methodology described here is fundamentally correlational. When experimental manipulations are performed on stimulus characteristics (i.e. lexical and/or semantic variables) and not on the cognitive processes presumed to be involved in the task directly, the results can only represent an indirect source of evidence for the format of conceptual representations. To return to our example, a stronger inference about the role of motor information could be made if, e.g., the variable's facilitatory effect disappeared in a second group

of participants whose hands were constrained to interfere with motor simulations (for a discussion of strong inference and causal paradigms, see Ostarek & Bottini, 2021. We will discuss such experiments in Chapter 2). We do not intend to imply in any way that this is a specific limitation in the word recognition literature. The same point can be made about many other experimental paradigms, and causal experiments come with their own set of challenges and limitations. This nevertheless implies that the results of such studies must be interpreted with caution and investigated at more length; they cannot provide any definitive evidence on representational format without more adequate experimental setups, however informative and valuable they may be.

A second and much more slippery challenge for deriving (2) is that it is inextricably dependent on (1), i.e. on the validity of our variables. If our manipulability scale does not reflect the extent to which motor information is associated with objects, or if it concurrently captures some other dimension in disguise, then the ratings can be misleading. The fundamental difficulty we are faced with here is that we have no systematic and reliable method to determine what our variables truly capture. The most recent editorial letter of *Behavior Research Methods* – the leading journal for normed materials – is quite telling:

> Validity refers to the fact that we are measuring what we claim to measure, which is crucial for the accurate interpretation of measurements. For research tools, this is often ascertained by correlating the measure with an external criterion. If, for instance, we see a new measure of word frequency, we want to know how well it predicts an important criterion, such as the processing speed of words. Otherwise, it is possible that the new norm measures something else (e.g., because a calculation error was made). Another way to collect evidence of validity is to compare the new measure with an existing measure (convergent validity). If the new measure is useful, it will correlate well with the existing measure, while improving on it in interesting ways. [...] Information about reliability and validity is central to BRM articles. (p.2, Brysbaert et al., 2020)

The above passage argues that a variable's validity should be determined with respect to its correlation with an external criterion and with other variables. These methods have indeed been the standard approach and can be informative. They are nevertheless also quite limited in determining whether our variables really capture "what we claim to measure". For one, assessing the predictive power of a variable on task performances (as suggested in the editorial) says fundamentally nothing about the variable itself and can even be misleading if, for instance, one argues that the effects of semantic variables differ across experimental tasks. Correlations with existing dimensions appear to be a safer bet and have been used extensively to gain insights into various dimensions and to detect potential confounds. However, this implies a rather cyclical process in which one variable's validity is determined against other variables which might themselves not be valid – at least in the sense that they might not be entirely reflecting what we presume. We will delve deeper into this

question in Chapter 3 with a concerning problem affecting a large number of semantic – and more generally Likert-type – variables, and see in Chapter 4 that it also unexpectedly brings a potential solution for the assessment of their validity.

1.2.5 Synthesis

We have seen that the visual word recognition literature has mostly converged with embodied cognition over the last years – and particularly with the hybrid views mentioned in Section 1.2.2. This line of research highlights a flexible cognitive system that draws on different processes to adaptively tackle specific situations. Tasks requiring a deeper processing of a word's meaning should involve richer sensory-motor simulations, while shallower lexical tasks could be solved by the linguistic system. Additionally, richer semantic representations are proposed to facilitate access to a word's meaning. In the case of manipulable objects and given a semantic task, the general hypothesis would thus be that processing their names will be faster and more accurate relative to those less strongly associated with motor experiences. In the next section we will briefly cover the visual processing of the objects themselves, and the insights gained from action and tool use research that will help us refine our understanding of 'manipulable objects'.

Before we continue, it is important to emphasize that the extensive investigation of widely accepted variables and tasks in psycholinguistics greatly facilitates cross-study comparisons and has produced an invaluable resource for methodological inquiry. While similar variables to those presented above have been used in other research areas, we will see in Chapter 2 that their implementation has often lacked the same rigor. This field will thus serve as our anchor as we start exploring the assessment of stimulus characteristics in Chapter 3.

1.3 A braid of action and perception

1.3.1 Vision: the what, the where and the how

Especially before the widespread use of neuroimaging techniques, mapping the functional organisation of the brain heavily relied on neurophysiological studies on non-human primates. In a seminal synthesis of their work and of the extant literature, Ungerleider and Mishkin (1982; Mishkin et al., 1983) put forth that visual information follows two separate pathways that originate in the primary visual cortex (V1). These *ventral* and *dorsal* streams (projecting from V1 to the temporal and parietal cortices respectively) were notably proposed to serve different functions. Through a series of meticulous neuroanatomical lesion and ablation studies, the authors found that

damage to the ventral stream produced severe deficits in discriminating and recognising visual patterns, but not in the ability to process spatial relationships between objects. Conversely, monkeys with a lesion along the dorsal pathway had normal visual recognition but drastically impaired performance in visuo-spatial tasks (e.g. determining to which of two positions an object is closest – the *landmark task*). Based on these results and on a rich body of related evidence, the authors concluded that "appreciation of an object's qualities and of its spatial location depends on the processing of different types of visual information" (p. 578). That is, the ventral pathway was proposed to be primarily concerned with information relevant to identifying objects (the *what*), and that the dorsal stream to handle spatial information (the *where*. For an earlier account of this distinction in rodents, see Schneider, 1969).

Although the segregation of the two cortical pathways was largely corroborated by subsequent studies, it also became increasingly apparent that the dorsal stream's function cannot be limited to spatial perception and that it extends to object-directed manual actions. For instance, Taira et al. (1990. See also Mountcastle et al., 1975) showed that some parietal cells (area 7 – dorsal stream) of Japanese monkeys fired when they manipulated different types of switches (e.g. pulling a lever or a knob, pushing a button). These cells were critically found to discharge when manipulations were carried out both in the light and in the dark (i.e. in the absence of visual information) and to be selective to the objects and to their orientations, but not so much to their position in space. Another significant source of evidence came from human neuropsychological case studies, and notably the famous case of patient D.F. who presented with visual form agnosia following damage to her ventral pathway (Goodale et al., 1991). Across a variety of tasks, D.F. had severely impaired perceptual judgements (e.g. to indicate the width of plaques with her thumb and index finger), but was indistinguishable from healthy participants in her visuo-motor abilities (e.g. to reach and grasp the plaques). The inverse pattern was also observed with lesions to the dorsal pathway, leading to a condition known as optic ataxia. Such patients typically retain their ability to recognise visually presented objects or patterns, but display systematic errors when reaching for them, as well as inadequate movement kinematics and finger positioning during prehension (e.g. Jackobson et al., 1991; Jeannerod, 1986; Jeannerod et al., 1994).

In their review of such evidence, Goodale and Milner (1992; Milner & Goodale, 1993) proposed a highly influential reinterpretation of the two visual pathways in terms of the computations they perform instead of the nature of visual information that they process. Somewhat similarly to Ungerleider and Mishkin's (1982) model, they proposed that the ventral pathway codes the invariable characteristics of objects to allow their identification and recognition (the *what*, or *vision-for-perception*). The dorsal pathway, on the other hand, was conceived as providing a description of objects with respect to the viewer's frame of reference (e.g. orientation, position) to allow the

on-line control of actions (the *how*, or *vision-for-action*). This reformulation continues to be the dominant paradigm in much research on vision and has fostered extensive empirical investigation both into the two streams and into their interactions (for a recent overview, see the special issue in *Cortex*, De Haan et al., 2018. For criticisms, see e.g., Rossetti et al., 2017). Most importantly for our current purposes, it provided the groundwork for bridging the fields of perception and action – to which we turn next.

1.3.2 The action in perception

While perception has historically been studied to a large extent in isolation – especially before the proposal of the 'how' pathway, early models of motor control were faced with a number of challenging problems that required the integration of perceptual elements. For instance, actions are generally carried out in order to attain a goal. This implies not only knowledge of how to move one's body given its current state (i.e. the complex coordination of muscles and joints. See e.g., Turvey et al., 1982), but also of the sensory consequences of the movements to both plan their desired outcomes and to monitor their execution. To account for these aspects, motor control has been generally conceived as a predictive and dynamic process that engages tightly integrated action and perceptual representations (e.g. Adams, 1971; Rosenblueth et al., 1943; Schmidt, 1975; von Holst, 1954). In a highly influential model, Schmidt (1975) notably proposed that actions are guided by 'generalized motor programs' (or *schemas*), i.e. 'templates' that are abstracted from the invariant characteristics of similar movement patterns and of their sensory outcomes. To take an example from the author, each instance of throwing a baseball would rely on the same schema, parametrised to align with the body's current state and the desired outcome. What was lacking, however, was their integration into a more general framework to understand cognition and behaviour.

Monkey see, monkey do

As with visual perception, most early insights about the neurophysiology of motor control came from the study of non-human primates. One of the major discoveries on the topic was made by Rizzolatti and his collaborators who investigated the role of an area in the premotor cortex of macaque monkeys (F5) during object-directed manual actions. In an early study using single neuron recordings, Rizzolatti et al. (1988) found that that the activity of F5 neurons was not correlated with specific movements as such, but rather with their goals. For instance, some neurons were active only when the monkeys brought food to their mouths but not during similar arm movement. Others were found to be specific to grasping actions and composed the majority of the neurons in

F5. Most of them were notably found to code specific grip types (i.e. precision grip, finger prehension, whole-hand grip). Among these, some were observed to be correlated with different phases of the action (e.g. finger extension, finger flexion), while others started firing slightly before any overt movements and until the object was grasped. Most interestingly, a subset of the 'grasping neurons' was also found to respond to the simple visual presentation of objects in the absence of any action requirements, and to be once again sensitive to object size and shape – i.e. to code for the specific grips that would be necessary to pick them up.

Subsequent studies confirmed these initial observations and further revealed the existence of two broader types of neurons within F5 that have both visual and motor properties, namely *mirror* and *canonical* neurons (e.g. Bonini et al., 2014; Di Pellegrino et al., 1992; Fadiga et al., 2000; Gallese et al., 1996; Kohler et al., 2002; Murata et al., 1997; Raos et al., 2006). Mirror neurons fired both when monkeys performed grasping actions and when they observed others performing them. This finding has notably led to highly influential accounts of social cognition by providing a potential neural basis for, e.g., imitation, empathy, and understanding the actions and intentions of others (Gallese, 2007; Iacoboni, 2009). On the other hand, canonical neurons were found to discharge during the execution of prehensile actions and the passive visual fixation of objects. In contrast to mirror neurons, the activity of this latter type of cells has been critically shown to be constrained to objects in peripersonal space and to be modulated by their orientation. Coupled with their selectivity for specific grip types, these neurons thus appeared to form a "vocabulary of motor acts" (p. 506, Rizzolatti et al., 1988) similar to the schema discussed above, and to "represent the description of [objects] in motor terms. That is, every time an object is presented, its visual features are automatically (regardless of any intention to move) "translated" into a potential motor action" [9] (p. 2229, Murata et al., 1997. For a discussion, see also Rizzolatti et al., 1998).

Simulated affordances

The discovery that neurons with a common sensory-motor code are sensitive to the mere presentation of objects strongly resonated with a slightly older and fringe concept: *affordances*. The term was coined by J. J. Gibson (1966, 1977, 1979) as one of the foundational ideas of

[9] It should be noted, however, that Rizzolatti et al.'s (1988) earlier study was more cautious about the automaticity of this activity. They emphasised that it is not consistently observed and that it was strongest when the objects were "motivationally meaningful" (p. 501), such as food. Subsequent studies claiming automatic activations used go/no-go experimental designs, where monkeys were cued with LED lights to either reach and grasp or fixate objects. To make a somewhat exaggerated analogy, using food stimuli in such an experiment is roughly equivalent to placing a marshmallow in front of a child and asking them to wait before eating it. Even with neutral stimuli (e.g. sphere, cylinder), the experimental setting implies the execution of an action that must be inhibited during no-go trials. It is thus reasonable to suspect that the discharge of canonical neurons was not a direct result of the passive fixation of objects, rather an effect of task demands and of the monkeys' intentions.

his ecological account of visual perception[10] (for the theoretical and philosophical foundations of this approach, see Lobo et al., 2018). Gibson was highly critical of the dominant computational-representational paradigm and of its experimental methods studying perception in isolation and as a passive process. In his view, we do not sense an objective physical world that is reconstructed (or represented) in our heads, to which we then ascribe value and meaning. He argued instead that the ambient energy in the environment (e.g. light, sound waves) is already structured and that organisms attuned to its regularities can directly pick up meaningful information through their movements, without any need for additional internal computations. Perception and action are here described as two sides of the same coin: we act in order to perceive ("We don't simply see, we look", p. 5, E. J. Gibson, 1988), and we directly perceive the possibilities for action that our environment offers – i.e. its affordances. Gibson defined affordances as relational properties between organisms and their environments, emphasizing that a description of the environment – and of the objects of perception – must necessarily include the observer and its biological capabilities. For instance, a mug affords grasping and drinking for most humans, but a (potentially dangerous) surface for the locomotion of small insects (for examples of experiments in ecological psychology, see Adolph, 1993; Oudejans et al., 1996; Warren, 1984).

Findings linking perception and action in the dorsal pathway provided a neural basis for affordances and strongly contributed to bringing the concept into the mainstream. Gibson's radical departure from the traditional paradigm – and perhaps his sometimes rather opaque writing – nevertheless led to a highly loose usage of the term and to its assimilation into a representational framework[11] (for discussions, see Chong & Proctor, 2020; Ferretti, 2021; de Wit et al., 2017; Declerck, 2013; Norman, 2002; Osiurak et al., 2017. For a tragicomic review of Gibson's misrepresentation in textbooks, see Costall & Morris, 2015). Declerck (2013) attributed the representational account of affordances to the influential *motor simulation theory* formulated by Jeannerod (1994, 2001, 2003, 2004). In a similar vein to multimodal simulations (Section 1.1.3), Jeannerod proposed that the motor system is capable of covertly simulating actions. He argued that these simulations (or representations) not only allow the anticipation and planning of actions, but that they are also a

[10] Note that his work was part of a joint project with his wife, Eleonore Gibson, who worked on perceptual learning and child development (e.g. Gibson, 1969, 2000. See also Adolph & Kretch, 2015).

[11] Although not specifically discussed here, let us note that Gibson's ecological approach has also been a major source of influence for more 'radical', non-representational accounts of cognition (see Section 1.1.3). One of the main difficulties in reconciling Gibson's views with the notion of representation has been the very definition of this term and the underlying assumptions about how cognition operates. Additionally, cognition is "representation-hungry" (p. 235, Golonka & Wilson, 2019) and it is not entirely clear how to dispense of representations entirely to account for the types of behaviour that we display (for discussions, see Golonka & Wilson, 2019; Norman, 2002). It has been proposed that the currently growing movement that describes cognition as a fundamentally predictive process might be able to reconcile the two views, providing "a way to end the representation wars" (p. 1, Constant et al., 2021. For the fundamental idea behind the predictive approach, I highly recommend the introduction in Clark, 2016, and Karl Friston's short interview in Serious Science, 2017).

means to gain "information about the feasibility and the meaning of potential actions" (p. 103, Jeannerod, 2001). Regarding objects specifically, Jeannerod (2003) noted that their representations "encode those properties [...] that are relevant to potential interactions with the agent, according to his intentions or needs" (p. 129). The use of the term 'affordance' in the cognitive psychology and neuroscience literatures has largely come to refer to this view, i.e. to the motor representation of the actions that can be performed with objects.

As Osiurak et al. (2017) have noted, "the notion of affordance raises serious theoretical issues, notably when the time comes to define precisely what it is" (p. 403). Indeed, the authors' review reveals that despite the general description of affordances as representations, there has been wide disagreements about what they refer to and about their neural implementation. Two critical and related points about affordances are particularly relevant for our current purposes: their automaticity, and the actions to which they refer. The question of whether we automatically 'simulate' potential motor interactions whenever we see objects and independently of the context was already hinted at in the quotes from Murata et al. (1997) and Jeannerod (2003). In the former case, the results from monkey neurophysiology were interpreted as an automatic activation of manual affordances irrespective of task requirements (but see footnote 9). On the other hand, Jeannerod (1994, 2001, 2003) has repeatedly highlighted the importance of goals and intentions, thus suggesting that manual interactions with perceived objects might not be systematically simulated. The still ongoing 'automaticity debate' was nevertheless mainly sparked by a series of influential – and controversial – behavioural studies by Tucker and Ellis (1998, 2001, 2004; Ellis & Tucker, 2000) which have been continuously cited in support of an automatic evocation of object affordances in the embodied cognition literature. We will cover this experimental paradigm in detail in Chapter 2 (Section 2.1). For the time being, let us only note that this question has prompted a rich inquiry into whether the visual perception of manipulable objects necessarily implies an activation of their affordances, and under which conditions they arise otherwise. In the next section, we turn to the second and highly related question: what affordances are.

The two routes of action

Most of the research discussed so far was focused on relatively simple object-directed manual actions such as pointing, reaching and grasping, while more complex actions were largely left unaddressed. The use of tools in particular cannot be reduced to visuomotor processing alone and must involve some extent of knowledge about how to use them, what they are used for, their mechanical properties, etc. However, how these aspects integrate with the two visual systems hypothesis (TVSH, dorsal: vision-for-action; ventral: vision-for-perception. Section 1.3.1) was not entirely clear. Should the learned use of tools, for instance, be considered as semantic knowledge processed

by the ventral pathway? Or are they represented in motor format along the dorsal stream (for a review of different views, see Osiurak et al., 2017)? Around the early 2000s, debates about this issue and several additional lines of evidence started converging on the idea that the TVSH requires further refinement (e.g. Buxbaum, 2001; Glover, 2004; Johnson-Frey, 2003, 2004, 2007; Pisella et al., 2006; Rizzolatti & Matelli, 2003; Tomassini et al., 2007). Two particularly influential arguments came from neuroanatomical studies in non-human primates and from human neuropsychological case studies. In the former case, syntheses of the extant findings and additional studies revealed that the dorsal pathway is composed of two parallel cortical circuits (see below). These roughly project from the visual cortex to separate premotor areas through different parietal structures, suggesting distinct functional specialisations (Rizzolatti & Matelli, 2003; Tanné-Gariépy et al., 2002. For a review and discussion, see also Johnson-Frey, 2003).

The second and most relevant evidence was presented by Buxbaum (2001) in her review and discussion of *ideomotor apraxia* [12]. Apraxic patients typically have no difficulty with the control of self-initiated basic manual actions towards objects (e.g. reaching and grasping, in contrast to optic ataxia, Section 1.3.1), nor in their declarative knowledge of what the objects are and what they are used for. However, they present significant impairments with more 'conceptual' actions. Although the deficits vary, they typically include errors when asked to pantomime the use of familiar objects (e.g. miming a hair combing action for the use of a toothbrush), to produce or recognise symbolic gestures (e.g. waving hello), and, in executing the appropriate grasp or movements for the use of real objects. Reviewing the different deficit patterns and drawing on the influential models of the time, Buxbaum (2001) reinterpreted the findings in light of the TVSH and argued that they can be accounted for by the existence of – and interaction between – three processing systems (see also Binkofski & Buxbaum, 2013; Buxbaum, 2017; Buxbaum & Kalénine, 2010. For similar views, see Borghi & Riggio, 2015; van Elk et al., 2014). The same *ventral system* as in the TVSH was proposed to store declarative knowledge about the function of objects, but not how they are manipulated (e.g. that 'knives are for cutting'). In line with the insights from primate neuroanatomy, the dorsal pathway was subdivided into two substreams: the *dorso-dorsal* (or Structure/Grasp system) and the *ventro-dorsal* (Function/Use system) pathways. The former was described as dynamically processing spatial and structural information (e.g. shape, size, orientation, position of objects) in the service of real-time action – i.e. similarly to the dorsal pathway's original formulation in the TVSH. The ventro-dorsal stream, on the other hand, was proposed to extract and represent the invariant features of actions across their different occurrences, i.e. to subserve learned actions (e.g. tool use, similar to the motor schema proposed by Schmidt, 1975).

[12] Different types of apraxias have been described in the literature (i.e. ideational, ideomotor, and limb apraxia). As Buxbaum (2001) has argued, however, there have been important disagreements about the criteria used to diagnose them, as well as the nature of the underlying deficits.

A wealth of recent neuroimaging studies have mapped a highly distributed and interconnected network of brain areas involved in tool use and recognition (sometimes called the *tool use* – or *processing – network*) that is generally consistent with the distinctions presented above (e.g., Bi et al., 2015; Brandi et al., 2014; Chen et al., 2023; Gallivan & Culham, 2015; Garcea & Buxbaum, 2019; Garcea et al., 2018; Mahon, 2020; Sakreida et al., 2016). However, let us also note that there is no general consensus regarding the network's dynamics, the specific role of its different structures, nor about the network's broader functional role in cognition. For instance, Fischer and Mahon (2022) have argued that the network is not specialised in tool processing as such, rather in generating predictions about "the future state of the world from a first-person perspective" (p. 464). There are additionally methodological concerns that cast doubts on the generalisability of most earlier neuroimaging findings. Freud et al. (2020), for instance, have raised concerns about the widespread use of 2-dimensional object images that not only lack information to which the dorso-dorsal stream can be sensitive to (e.g. contextual cues, stereoscopic depth), but are also typically presented in unrealistic sizes – an observation that can be extended to most behavioural studies in general. These disagreements and issues aside, this literature's rich history suggests that visual tool processing (e.g. perception, recognition, naming, miming, using) involves distributed and multimodal neural structures that are highly dynamic and sensitive to both the agent's goals and to the situation at hand.

1.3.3 Open questions about manipulable objects

The model presented in the previous section essentially describes two types of motor processes that must necessarily coordinate during our interactions with manipulable objects. The first monitors the execution of movements, is highly sensitive to bottom-up visual information and appears to describe visually seen objects in terms of how they can be 'structurally manipulated' (i.e. reached and grasped). The second one subserves the representation of learned actions, and thus the 'functional manipulations' associated with objects (i.e. the movements necessary for their use). Taking a step back, this implies that only functional actions are a type of knowledge, structural actions being largely transient and dependent on the immediate interaction. For the sake of simplicity, in what follows we will use 'manipulable object' (MO) to refer to objects that both allow structural manipulations (SM) and are associated with a typical functional manipulation for their use (FM).

In light of the distinction between motor processes and the discussions surrounding the grounded representation of knowledge (Section 1.1.3), the 'automaticity debate' mentioned earlier can be reframed as two questions that have mutual implications: does merely seeing MOs automatically evoke their SM? And, does conceptual processing of MOs necessarily involve simulations of FM? The debate on both of these questions is very much open and weighing the evidence for each

– especially for FM – will be the main objective of Chapter 2. Before we delve into them, there are nevertheless several methodological constraints that we must bear in mind for a careful consideration of the extant literature, which have also been aptly highlighted by Buxbaum and Kalénine (2010) and Kalénine and Buxbaum (2015).

First, if it is the case that seeing a MO automatically activates SM, then any conclusion about an activation of FM during object processing would be equivocal if the study's methodology does not clearly distinguish the two. Conversely, if recognising a MO recruits FM, this could lead to the false inference of SM being automatically evoked. These points alone make this topic highly challenging to study because objects associated with FM are often also associated with SM, and vice versa. An additional and critical consideration that is not consistently addressed by studies investigating the role of motor information in object processing is how motor processes interact with perceptual and attentional mechanisms. Indeed, we have mentioned that motor representations are highly integrated with sensory information[13]. Motor activity can thus involve expectations about its sensory consequences and draw attention towards relevant information for proper planning and control. In line with this, there is ample evidence that preparing a manual action biases attention towards – and facilitates the processing of – compatible stimuli and visual features (e.g. Craighero et al., 1999; Fagioli et al., 2007; Hannus et al., 2005; Wykowska et al., 2009. See also Section 2.1). As the findings regarding mirror neurons also suggest, seeing others perform manual actions, or simply pictures of hand postures, can similarly prime the motor system and result in the same effects (e.g. Bub et al., 2021; Vogt et al., 2003). We will return to these points in Chapter 2. Without adequate methodological safeguards, studies requiring manual actions (especially grasping) or using hand (or action) representations thus introduce task-induced confounds and are not readily informative about the automaticity of either SM or FM. Finally, a large number of studies investigating motor effects in object processing have required participants to provide keypress responses. Despite the fact that these imply a constant engagement of the hands and entirely different movement mechanics than grasping and manipulating an object, the literature is, to our knowledge, largely silent on its potential effects on the evocation of either SM or FM.

[13] We have only alluded to this point in the description of neurons that discharge to both movement and sensory information. There is nevertheless a rich literature on the integration of motor and sensory information, notably in the general context of *ideomotor theory* (for a review, see Shin et al., 2010).

1.3.4 Synthesis

The research presented in this section has largely shaped our current understanding of the functional organisation of visual object processing and tool use. It notably highlights a tight coupling between perception and action, suggesting that contextual elements – and even visual properties – can readily prime motor representations. This point is crucial to consider in experimental settings as it could otherwise obscure the role of learned, functional manipulations in the conceptual representation of objects.

Armed with these methodological constraints, and with the overall theoretical and methodological framework for the study of conceptual representations, we can finally turn to the empirical evidence on whether object knowledge is (partly) grounded in motor information.

Chapter 2

Evidence for the role of motor representations in object processing

Our goal in this chapter will be to assess the empirical evidence for the activation and functional role of actions during object processing. In light of the views presented in Chapter 1 about the flexibility of the conceptual system (Sections 1.1.4 and 1.2.2), we will try to cover a broad range of experimental paradigms to determine whether some effects are more reliably observed in specific tasks. Given the scope of this review and thesis, we will nevertheless limit ourselves to behavioural experiments and thus mostly omit neuroimaging studies. The first reason for this is that a faithful discussion of the latter would require technical details and vocabulary which have not been previously presented, adding significant complexity to an already dense content. The second reason is that neuroimaging findings are fundamentally correlational and do not significantly add to the debate. As Pecher (2013b. See also Fischer & Mahon, 2021; Poldrack, 2011) notes:

> When researchers observe brain activation in overlapping areas between two tasks, they tend to conclude that the tasks must share a functional component. Such a conclusion is only valid, however, if brain regions are involved in only one particular function, and we know this is not the case. When researchers still draw conclusions about function from activation in certain brain areas, they are committing the logical fallacy of affirming the consequent. (p. 11)

We will be particularly attentive to the role of functional manipulations (FM) as these are typically considered to be constitutive of conceptual representations, whereas structural manipulations (SM) are presumed to be mainly driven by bottom-up visual information (Section 1.3.2). As pointed out in Section 1.3.3, however, gaining a better understanding of the conditions under which SM are evoked will also be essential for more reliably assessing the evidence for an involvement of FM. To that end, we will start the chapter with a paradigm that mostly assesses the automaticity of SM (Section 2.1), and then present experiments that should arguably draw more heavily on conceptual knowledge and thus FM (Sections 2.2 through 2.4). While our focus will primarily be on methodological considerations, it is important to keep in mind that many of the studies discussed below suffer from small sample sizes and sometimes overly complex statistical designs. It is thus overall difficult to determine the extent to which both significant and null findings are reliable.

2.1 Affordances and compatibility effects

Claims such as “perception of objects automatically evokes potential actions to interact with those objects” (p. 1, Zhao, 2020) are common across the literatures of embodied cognition and object processing. These generally draw on the notion of affordances (Section 1.3.2) and on a line of inquiry pioneered by Tucker and Ellis (1998). As we will explore in this section, however, experimental effects attributed to an automatic (i.e. goal- and context-independent) activation of object affordances have been – and continue to be – highly debated.

The experiments discussed in this section have used a stimulus-response compatibility (SRC) paradigm. As its name suggests, this paradigm is used to investigate how reaction times vary as a function of the congruence between a stimulus and the required response, with faster responses occurring when the two are compatible (vs. incompatible). In the current context, the stimuli generally consist of graspable objects and participants are asked to make a manual response that is either compatible or incompatible with how they can be grasped (e.g. responding with the hand that is aligned with, or opposite to, the object's handle). Critically, the stimuli's feature whose compatibility with the responses is being investigated in an SRC paradigm is task-irrelevant to ensure that any observed effect arises automatically from the visual presentation of the stimuli (e.g. judging the colour of objects that have handles pointing in different directions). Thus, findings that responses are facilitated when they are compatible with how the objects can be grasped can be interpreted as evidence that the objects automatically elicit their affordances – i.e. how they can be acted on. In the following sections, we will present brief reviews of two specific methodologies used to investigate this question, namely studies using keypress responses, and others asking participants to execute grasps. For simplicity's sake, we will use the term *affordances* and *manipulation* in this section to refer to any type of manual action that can be performed with objects, as structural and functional manipulations have typically not been distinguished in this literature and because they will be more specifically reviewed later in this chapter.

To provide some context on the debates, alternative views to the affordance account explain compatibility effects based on abstract codes and attentional mechanisms. A particularly relevant example is the Simon effect (1969, 1990), in which reaction times are influenced by the correspondence between the location of a stimulus – or even the location that it implies – (e.g. left or right) and the relative side of the response (e.g. left or right hand). Similar and more abstract mappings could also occur with other stimulus characteristics, such as their size or shape (Kornblum et al., 1990). Several authors have thus argued that findings of a compatibility between object affordances and manual responses do not readily provide strong evidence for this account (e.g. Azaad et al. 2019; Heurley et al., 2020; Proctor & Miles, 2014).

2.1.1 Keypress responses

In a seminal study, Tucker and Ellis (1998) asked participants to determine whether pictures of individual manipulable objects (e.g. *frying pan*, *saw*, *teapot*) were upright or inverted, by giving keypress responses using their left and right hands (e.g., left for 'upright', right for 'inverted'). The handles of the objects were oriented either to the left or to the right but were irrelevant to the task. The hypothesis of the authors was that the orientation of the handle would evoke motor representations with the hand that is on the same side. Thus, irrespective of the upright/inverted

presentation of the objects, responses were expected to be fastest when the handle and the responding hand were aligned, and slower when they were incompatible. Their results were as predicted, with a significant interaction between the responding hand and the handle's orientation (we will refer to this as the *compatibility effect*). However, the same finding could be explained by a spatial correspondence, i.e. due to the handles drawing attention to the side of the response (the Simon effect). To rule out this possibility, Tucker and Ellis (1998) repeated the experiment, this time with participants responding with their index and middle fingers of the same hand. They argued that if the initial effect was due to a mapping between the response option and the handle's spatial location, the same compatibility effect should be observed under this version of the experiment as well. Instead, the results revealed no compatibility effect, leading to the conclusion that the first experiment's findings were driven by an automatic activation of object affordances.

Subsequent studies, however, showed that this initial claim was not as straightforward and generalisable as it appeared. For instance, Phillips and Ward (2002) found the same compatibility effect when participants responded with their hands crossed (but see Janyan & Slavcheva, 2012) and when they responded with their feet (see also Symes et al., 2005), thus challenging the effect's sole attribution to affordances. Other studies focused on the task's decision requirements and reported that the compatibility effect is observed when participants judge the shape (round, square), the orientation (upright/inverted) or the semantic category of objects (kitchen/garage), but not their colour (Pellicano et al., 2010; Saccone et al., 2016; Tipper et al., 2006. But see Cho & Proctor, 2010). These results appear to show that the evocation of affordances is contingent on processing the perceptual or semantic characteristics of the objects. However, Tipper et al. (2006) noted that judgements about object shape yielded reliable compatibility effects only when participants were previously shown videos of people manipulating similar objects. In contrast, Cho and Proctor (2013) failed to replicate Tipper et al.'s (2006) results, both with and without a video. Yet another study by Yu et al. (2014) failed to replicate Tucker and Ellis's (1998) initial findings (upright/inverted judgements. See also Matheson et al., 2014a), and only found a compatibility effect when participants were instructed to imagine picking the objects up during the task. The evidence for how decision requirements modulate the compatibility effect thus remains largely controversial. There appears to be some indication that explicitly drawing attention to the manipulation of objects can result in the activation of affordances. Note, however, that this result could once again be explained by a spatial correspondence effect. In a replication attempt of Yu et al. (2014), Thomas et al. (2019) argued that imagining picking the objects up might simply bias attention towards their graspable part and facilitate responses on the corresponding side. Indeed, they showed that the same compatibility effect was found when participants responded with two fingers of the same hand, as well as with their feet. The conclusion that affordances are evoked by processing how an

object can be manipulated is thus also disputed.

An issue that significantly complicates the interpretation of these results is that the compatibility effect appears to be extremely sensitive to how the stimuli are positioned on the screen. Bub et al. (2018. See also Proctor et al., 2017) notably showed that a compatibility effect was obtained for upright/inverted judgements when the objects were centred so that an equal number of pixels appear on both sides. When the centre of the entire object was used, however, the compatibility effect reversed entirely and responses were fastest when the handles were oriented opposite to the responding hand[14]. This point has received relatively little attention, but how objects are centred – and their asymmetry – appears to significantly affect compatibility effects and to yield different results depending on other methodological choices (for discussions, see Bub et al., 2021; Masson, 2018).

A recent meta-analysis by Azaad et al. (2019) sought to determine the overall compatibility effect and to assess how different moderators such as those mentioned above (e.g. decision type, response type, centrality) affect the results. Over 36 publications and 88 independent effects, the authors found a (very) small effect size, in addition to significant publication bias and between-study heterogeneity. Following their moderator analysis, they further concluded that their results mostly support a "spatial account of *object-based CE*" (p. 117, emphasis in original, CE: compatibility effect). It is important to stress, however, that given the differences in methodologies and the sensitivity of the effect to small variations (and to their interactions), it is highly possible that the specific contexts that influence the compatibility effect might have been obscured by the meta-analysis (for a similar point, see Bub et al., 2021).

There are also a number of other limitations that raise concerns about the interpretability of the available findings. First, a large number of experiments have used only a very small number of stimuli (sometimes a single object, e.g. Pappas, 2014) that are repeated across many trials. Whether this might affect the activation of affordances, however, is largely unknown. Taking a context-dependent approach to affordances, one could reasonably suspect that a 'surprising' stimulus evokes potential motor interactions that attenuate with subsequent presentations. Second, the effects obtained in this literature are roughly of the order of 10 ms in most cases. While some early studies might have been statistically underpowered (e.g. there were respectively 8, 8 and 13 participants in the three experiments reported by Phillips & Ward, 2002), recent studies generally

[14]Centring an object such as a frying pan relative to its pixel count (as well as to the width of its base) results in the base staying relatively stable across trials while the handle strongly protrudes to one side. The stimulus thus draws attention to the side of its handle. In contrast, centring on the entire object's width leads to both the base and the handle to change considerably in position from trial to trial. For highly asymmetrical objects (e.g. frying pan), this also results in the relatively larger base to protrude on the side opposite to the handle. Bub et al. (2018, 2021) argued that the reverse compatibility effect is due to the object's base biasing attention towards the side contralateral to the handle.

perform power analyses to determine their sample sizes. These analyses, however, are far from indicative of an experiment's precision (e.g. Kelley et al., 2003; Trafimow, 2018) and, most importantly, do not account for measurement error. Standard commercial keyboards such as those likely used by a large number of studies have latencies that can range anywhere from 10 ms up to 50 ms on average, along with having large latency variances (Luu, 2017 ; Neath et al., 2011; Wimmer et al., 2019). When other sources of variability such as monitor refresh rate, software, driver and operating system delays are added, it is not at all evident whether the reported findings – or their lack – are genuine, or simply a result of hardware noise. Finally, the fundamental premise that object-evoked motor representations facilitate keypress responses is in itself questionable. As Suzuki et al. (2012) noted, "the action most strongly afforded by graspable objects is a grasping action, not a key press" (p. 882). A reach-and-grasp action and a keypress response are entirely different in terms of their kinematic and postural characteristics, their goals and their sensory consequences. Additionally, participants in most experiments keep their hands on the keys and thus execute only very limited movements. The only shared feature between the two actions is thus reduced to the effector being used, and Tuck and Ellis (1998) gave no particular justification on this point other than "one might expect [the preferential activation of the hand most suited to perform a reach-and-grasp movement] to facilitate simple keypress responses carried out by the congruent hand and, conversely, to interfere with those same responses carried out by the incongruent hand" (p. 833). Given the differences between the two actions, one might as easily argue that an object's affordances would interfere with keypress responses of the same hand or to have no effect at all.

Overall, the results from keypress studies thus appear extremely difficult to interpret. The current literature makes a convincing case for the compatibility effect with such responses to be partly due to spatial codes, while the questions of whether affordances are involved and under what circumstances are not entirely clear. If anything, this line of studies shows how sensitive experimental effects can be to small methodological differences and how easily one can "mistake an attentional effect based on spatial correspondence for evocation of a limb-specific action representation" (p. 223, Masson, 2018). As we will see next, however, important insights have been gained from studies using a different response modality, namely grasping.

2.1.2 Grasp responses

Ellis and Tucker (2000) investigated the potentiation of affordances through an alternative protocol involving a hand-held response device. Participants could respond by pressing either a small electrical switch between their index finger and thumb to mimic a precision grip, or a cylinder held in the same hand to imitate a power grip. Instead of the orientation of handles as with the experiments discussed above, the critical manipulation in this experiment was the size of the

objects and their congruence with the two grips (i.e. small – precision grip, large – power grip). Participants started each trial by seeing an object and were then presented with either a high or a low-pitched sound to cue one of the manual responses. The size of the object was thus irrelevant to the task. As the affordance account would predict, they found an interaction effect between the response type and the grip afforded by the objects. That is, participants were faster to respond with a precision rather than a power grip to small objects, and vice versa for larger objects[15]. Using the same response device, Tucker and Ellis (2001) found a similar grip compatibility effect when participants performed a semantic categorisation (natural/manufactured) with small and large objects. Grèzes et al. (2003) further reported a replication of this result in an fMRI study. They notably showed a correlation between dorsal pathway activity and the difference in reaction times between incongruent (e.g. small object – power grip) and congruent trials (e.g. small object – precision grip), which they attributed to a competition between manual responses and the affordances evoked by the objects. Although these results suggest that seeing objects automatically potentiates their affordances in the dorsal pathway, a series of experiments by Tucker and Ellis (2004) forced a reinterpretation of this position. Indeed, they found that the grip compatibility effect was present even when participants responded to objects that were occluded, degraded, masked or entirely absent from view. Additionally, the effect was found when object names were used as stimuli, overall suggesting that the presence of the objects while responding is irrelevant. The authors thus concluded that their "emphasis on dorsal system processing was probably misplaced" (p. 200) and that affordances can be evoked both from visual properties and from stored knowledge.

While this interpretation remains possible, Tucker and Ellis's (2004) results also raise the question of whether the effects can be attributed to affordances altogether. Indeed, Proctor and Miles (2014) have argued that "[s]ize is simply another dimension that yields SRC effects" (p. 254) and that the results can be simply due to a mapping between the size of the objects and of the different grips (Heurley et al., 2020; Kornblum et al., 1990). They additionally noted that the network identified by Grèzes et al. (2003) was found to be involved in spatial SRC studies and thus that this result does not necessarily imply affordance processing (e.g. Cieslik et al., 2010; Liu et al., 2004). The studies reporting grip compatibility effects also contain some inconsistencies. For instance, Grèzes et al. (2003) who attempted to replicate Tucker and Ellis's (2001) findings with a semantic categorisation noted that "motor responses were fastest to compatible objects and slowest to incompatible ones, showing that object affordances were extracted during the categorization task despite being task-irrelevant" (p. 2738). A closer look at their results nevertheless reveals that this was not the case. Only power-grip responses were faster to large (vs. small) objects, while

[15]Note, however, that this effect was only found when the precision grip was associated with a high-pitched sound (and the power grip to a low one) and absend in the reverse mapping, which "complicates the account" (p. 460, Ellis & Tucker, 2000).

precision-grip responses were even slightly longer to small objects compared to large ones. Similar patterns are present throughout the experiments conducted by Tucker and Ellis (2001, 2004). A significant interaction is generally present, but only one of the compatible pairings yields the expected effect while the other remains unchanged (e.g. no difference in performance for precision-grip responses to small and large objects, and a difference for power-grip responses). If affordances are indeed potentiated by the visual presentation of objects or by their knowledge, it is not clear why their effects would manifest only in some cases and not in others.

As Bub, Masson and their collaborators have repeatedly stressed, a major issue in this literature is that "[t]he role of intentional set has largely been ignored in the debate on whether limb-specific representations are potentiated by task-irrelevant objects" (p. 79, Bub et al., 2021). These authors arguably produced the most cautious and meticulous set of studies on how contextual demands and particularly the agent's action intentions contribute to compatibility effects. For our current purposes, one of their main assertions has been that objects do not automatically evoke the actions that can be performed with them. Rather, it is the preparation to perform an action that leads to an automatic evaluation of its outcomes, modulating attention towards features of the environment that are relevant to it (see also Section 1.3.3). Bub and Masson (2010) notably conducted a series of experiments in which manipulable objects in two colours were presented with their handles oriented to the left or to the right – similar to those discussed above. Participants were instructed to respond to the object's colour by reaching and grasping either a congruent or an incongruent 3-dimensional shape. The authors found a significant compatibility effect, with faster reaction times when the responding hand was aligned with the object's handle and the grasp was congruent with the object. However, a second experiment requiring participants to give keypress responses revealed no compatibility effect. Bub et al. (2008) similarly showed that the effect found with reach-and-grasp responses to the object's colour disappeared when participants were instead asked to reach and only touch the response device. More recently, Bub et al. (2021) extended the results and showed that compatibility effects attributable to object affordances emerged only when participants performed a reach-and-grasp action, or when they made keypress responses to pictures of hands (in a first-person perspective) superimposed on the objects (see also Bub et al., 2018; Ferguson et al., 2021; Girardi et al., 2010). Reviewing several lines of evidence and details of their experiments that go beyond our scope, the authors concluded that objects elicit motor representations only when a reach-and-grasp action has to be planned – either voluntarily or primed by the presentation of hand pictures depicting a grasping posture (i.e. implying a goal).

2.1.3 Synthesis

Several conclusions can be drawn from the studies discussed in this section. First, the SRC paradigm with keypress responses appears to be too vulnerable to confounding effects and does not allow to reliably determine if seeing objects automatically evokes the actions that can be performed with them. However, the evidence from studies using reach-and-grasp responses (especially those by Bub, Masson, et al.) strongly suggests that either preparing an action that is compatible with how an object can be manipulated or seeing the object in such a context (e.g. with hand postures) is necessary to evoke motor representations. It thus appears that it is rather the intention to act more than the object's intrinsic representation that potentiates actions. In the remainder of this chapter, we will be particularly attentive to how participants were required to respond across the different studies in order to better assess the evidence for whether a given type of task necessarily leads to motor simulations.

2.2 Object processing

Whereas the studies discussed in the previous section mostly focused on whether visual information can evoke motor representations (i.e. structural manipulations), others sought to determine if they are involved in conceptual processing. This question has been investigated through a number of different tasks that are presented in separate sub-sections below. Note that the studies in this section have typically used more specific definitions of manipulability, distinguishing the functional from the structural manipulations associated with objects (SM and FM respectively). This has been sometimes achieved through manipulability ratings that are presumed to assess these different dimensions. We will nevertheless see that their interpretation is not always straightforward (see also Chapter 4).

2.2.1 Identification

In light of a potential role of sensory-motor information in object recognition, Magnié et al. (2003) were among the first to have assessed object manipulability to allow the proper control of experimental stimuli. Participants were asked to rate line drawings of objects (Snodgrass and Vanderwart, 1980) on how easily they could mime the action associated with them so that someone else would be able to recognise the object (*pantomime* ratings). Thus, manipulability was operationally defined as the distinctiveness of the actions typically performed with objects and was proposed to mainly capture functional manipulations (FM). In a second study, they conducted an object decision task to determine how manipulability and other factors affect object recognition

performances. Participants were presented with black and white line drawings and had to decide if they depicted a real object or not[16] through keypress responses. The analysis showed that higher manipulability ratings were associated with both shorter decision latencies and higher accuracy rates after controlling for critical variables such as familiarity and visual complexity. This study thus suggested that manipulability – and functional use in particular – plays a role in object identification. Note, however, that manipulable objects (MO) are typically acquired early in life, which might have influenced performances but was not controlled.

2.2.2 Name matching

A number of other studies have investigated the role of motor representations in object recognition through tasks requiring to match objects to their names. One such task is the visual world paradigm in which participants are presented with multiple objects on a display while their eye movements are recorded. They then hear the name of an object and have to identify the corresponding picture ('target'). Among the other presented items, one typically shares a critical feature with the target (e.g. manner of manual use; 'competitor'), while the others are unrelated. If the investigated attribute is involved during object recognition, then an attentional competition between the target and the competitor would be expected, and consequently higher proportions of eye fixations to competitors compared to unrelated items. In a first such experiment, Myung et al. (2006) used competitor objects that either shared similar FM with the target (e.g. *piano – typewriter*) or had a similar shape (e.g. *piano – couch*; control condition). They found that participants made slightly more fixations to competitors in the former case, while there was no difference between competitors and unrelated items in the control condition. In support of multimodal simulations in conceptual processing, they concluded that FM were automatically activated by the object's name. Lee et al. (2013) extended this result by investigating the competition generated by objects that either shared similar structural manipulations (i.e. that can be grasped in a similar way; SM) but not FM (experiment 1), or similar FM but not SM (experiment 2). The object names were additionally associated with different verbal contexts. In half of the trials, the names were preceded by a neutral verb (e.g. *she saw the* ...), and in the other half by an action verb (experiment 1: *she picked up the* ...; experiment 2: *she used the* ...). They found that both types of manipulations created competition with the target, but with different time courses. SM competitors were fixated more than unrelated items in the early phases after an object name's onset. In contrast, FM competitors attracted more fixations relatively later and on a more sustained period. The verbal context was also

[16]The 'not real' category corresponded to chimeric objects that were constructed by combining one half of two different objects.

found to modulate the pattern of fixations, with the action verb appearing to drive the competition relatively earlier for both SM and FM distractors.

The experiments presented above bring elegant evidence for the activation and time course of both SM and FM, as well as for their modulation by the task's context during object processing. There are nevertheless some methodological points that require a more cautious reading of their findings. First, neither experiment controlled a highly critical dimension that could have affected the results, namely the visual complexity of the objects. For instance, the stimulus list provided by Myung et al. (2006) suggests that competitors (e.g. *piano*, *baby carriage*, *camera*, *corkscrew*) were generally more complex than the unrelated ones (e.g. respectively, *blanket*, *safety pin*, *needle*, *tooth pick*), which could have easily biased attention towards the former. Lee et al. (2013) did not provide their full stimuli, but the same concern applies. Myung et al.'s (2006) results are also rather counterintuitive in light of multimodal simulations. If processing an object's name involves a simulation of its perceptual and motor characteristics, then a competition with similarly shaped objects would have also been expected – not only with similarly manipulated ones. Finally, both studies required manual responses whose influence on the presumed conceptual motor representations largely remains unknown. Lee et al. (2013) notably required mouse-click responses, which implies that participants were continuously manipulating the mouse during the trials (Myung et al. asked participants to reach and touch the target picture on the display). It is not unreasonable to suspect that correspondence effects with the stimuli might have emerged as a consequence if object recognition does indeed involve motor representations. Keeping these points in mind, the results of these studies remain highly interesting, with Lee et al.'s (2013) results particularly pointing to a dissociation between SM and FM – although further investigations would be required to confirm the effects.

A few studies have also used tasks in which participants determined whether a noun matched the picture of a single object through keypress responses. Helbig et al. (2010) presented participants with videos of hands performing FM with invisible objects. A picture of a FM object that was either congruent or incongruent with the video was then very briefly shown, followed by the name of an object that could match or mismatch the picture. Participants were more accurate in their responses for congruent than incongruent trials – although not faster, thus suggesting that FM played a role in recognising the objects. Note, however, that the congruent primes could have simply drawn attention to the action-relevant features of the objects. The lack of an effect on latencies additionally implies that lexical-semantic processing was not affected, which would have been expected if FM were part of conceptual representations. In one of their experiments, Matheson et al. (2018) presented object names first, followed by object pictures. The manipulability of objects was studied as two continuous variables referring to SM and FM respectively (higher values

representing higher manipulability, Salmon et al., 2010), and the authors hypothesised that both types of information would facilitate object recognition. As SM are largely thought to be driven by bottom-up information, a visual mask was also superimposed on half of the images in order to interfere with the activation of SM. They found that higher SM values were associated with faster reaction times (but not higher accuracies) for unmasked compared to masked stimuli. However, FM had no effect on performances irrespective of the mask's presence, despite its presumed role in conceptual processing. It is possible that the task's requirements led participants to preferentially focus on visual characteristics instead of actions as object names had to be matched to pictures. On the other hand, the definitions used for SM and FM are rather counterintuitive. SM was defined as the ease with which objects could be grasped *and* used, and FM as the extent to which the movements to grasp an object differed from those to use it (Salmon et al., 2010). The definition of SM thus included functional actions, and whether the FM dimension reliably captured functional use is questionable (we will return to these questions in Chapter 4). It is thus difficult to determine what this study's results truly reflect.

2.2.3 Naming

The majority of experiments on the role of motor representations in object recognition have used variations of more straightforward object naming tasks. Salmon et al. (2014) asked participants to name black and white pictures and line drawings of the same SM and non-SM objects (SM: grasp and use, Salmon et al., 2010). They found that only pictures of SM objects were processed faster than line drawings (but not more accurately), and concluded that the richer visual cues in pictures facilitated their recognition by engaging motor representations. In a subsequent study with masked and unmasked pictures and continuous variables, the same team found that SM facilitated naming only for unmasked objects. Conversely, FM facilitated the recognition of only masked items (only for reaction times in both cases, Matheson et al., 2018). They thus argued for a dissociation of SM and FM, and that the presence of adequate visual information modulated their roles in recognition. Note, once again, that what the SM and FM dimensions in both of these studies capture is not entirely clear (see above). In contrast, Guérard et al. (2015) conducted a regression analysis on object picture naming latencies and found that both SM and FM facilitated performances above other critical variables (e.g. visual complexity, familiarity – although not the age of acquisition). More precisely, they used continuous manipulability ratings with 2 SM variables (the ease to grasp, and the ease to move objects) and 2 FM variables (ease to pantomime[17]

[17]This variable is similar to the one introduced by Magnié et al. (2003) mentioned earlier. The instructions were nevertheless slightly modified so that only the ease to mime the typical actions was assessed, irrespective of how easily they could be recognised.

and the number of actions associated with objects). Both FM variables and the ease to grasp objects were associated with shorter latencies. Objects that were easier to move were nevertheless named more slowly, which they interpreted as these items being more difficult to grasp and associated with less motor information. They thus concluded that both FM and SM facilitate object recognition.

McNair and Harris (2012) used an object-object priming paradigm and orthogonally manipulated the similarity of SM (grasp) and FM (use) between the object pairs (i.e. same SM and FM, same SM and different FM, same FM and different SM, and different SM and FM). Objects with similar SM were named more accurately than those with dissimilar SM, regardless of the correspondence of FM between the pairs. They thus concluded that only SM were involved in object recognition. Helbig et al. (2006) and Kithu et al. (2021) similarly used object-object priming but specifically focused on object pairs that were associated with similar or different movements for their use (FM). Helbig et al. (2006) reported finding "superior naming accuracy for object pairs with congruent as compared to incongruent motor interactions" (p. 221). Interestingly, the effect was no longer present when the prime pictures were replaced by the name of the objects. They thus concluded that FM facilitate object recognition but that they rely on visually derived information. It should be noted, however, that their analysis on picture primes did not reach significance when the stimuli were entered as a random variable to the statistical model. Conversely, Kithu et al. (2021) found no evidence that the congruence of FM between prime and target affected naming performances, whether the primes were 2-dimensional images or real objects. Finally, Ni et al. (2019) used hand primes (static or videos) instead of objects. These represented SM and FM that could be either congruent or incongruent with the ones required to manipulate the target objects. Across 4 experiments, they consistently found that only the congruence of FM between the primes and targets affected naming latencies. SM and FM were unfortunately not clearly defined, but the study suggests that SM corresponded to the possibility of grasping the objects and holding them in hand, while FM reflected functional use.

The results presented above make it difficult to draw general conclusions about the role of motor representations in object naming as they both reveal variable – and discrepant – effects of SM and FM on performances, as well as inconsistent findings across latencies and accuracies. These studies generally contain important confounds such as the failure to control critical variables that could have driven the effects, or the use of hand representations that could have biased attention towards action-relevant stimulus characteristics. Several additional methodological issues further complicate the interpretation of the findings. First, we have alluded to the fact that manipulability has been defined in different ways across studies, making the comparison of their findings difficult (see also Chapter 4). The practical statistical usage of continuous manipulability variables also raises some concerns as manipulability-related ratings are necessarily correlated but are generally

used concurrently without further control. For instance, Guérard et al. (2015) found opposite effects on naming latencies for the 'ease to grasp' and the 'ease to move'. The two variables were nevertheless entered in the same statistical model despite being extremely correlated ($r = .973$). This can severely impact the results (Friedman & Wall, 2005; Hsu & Chiang, 2020; Nickels et al., 2022), raising questions about the validity of the findings. Overall, the evidence from the above studies thus appears largely inconclusive.

A few studies have sought to provide causal evidence for the role of motor information in object recognition by using manual interference tasks. In the first study of this kind, Witt et al. (2010) used a task inspired by the SRC paradigm described in Section 2.1. Participants were presented with pictures of tools and animals in two orientations along the vertical axis and were asked to name them. In half of the trials, they additionally applied constant pressure on a foam ball with one of their hands to determine if interfering with motor processes affects object recognition. The authors found that tools were named faster when their handles were oriented towards the unoccupied hand than when they faced the occupied hand, whereas the orientation of animals did not affect performances. With shorter stimulus presentation times, the effect was also found for naming accuracies. These results have been highly cited as strong evidence for a role of the motor system in object recognition. Noting some limitations with their analyses, however, Witt et al. (2020) reanalysed their original data with more robust statistical methods and instead found substantial evidence for the null hypothesis (i.e. the concurrent manual task did not influence performances for tools) and concluded that "the current data do not advance [the debate over whether motor processes can inform cognition] as had been originally argued" (p. 1038). Additionally, two replication attempts of the initial findings failed to provide any evidence for a motor interference effect on naming tools, even with more dynamic manual tasks (Matheson et al., 2014b; Saccone et al., 2021). Another highly cited study was conducted by Yee et al. (2013) and investigated the interfering effect of a concurrent manual task relative to participant's experience interacting with the objects. In half of the trials, participants named the objects while repeating a sequence of hand positions with both of their hands. Partly as predicted, the authors found that the manual task interfered more with naming accuracies – but not latencies – when participants had more experience manually interacting with the objects than when they had less experience. There are nevertheless several issues with this study that do not readily allow to draw a conclusion about the role of motor information in object recognition. First, if the manual task interfered with conceptual processing, it should arguably have affected naming latencies as well. Second, manipulation experience was defined as the amount of experience participants had *touching* the objects, which can be quite different from their experience manipulating them. Critically, the experience ratings were not provided by the same participants who took part in the naming study and thus did not actually reflect their experience with the objects.

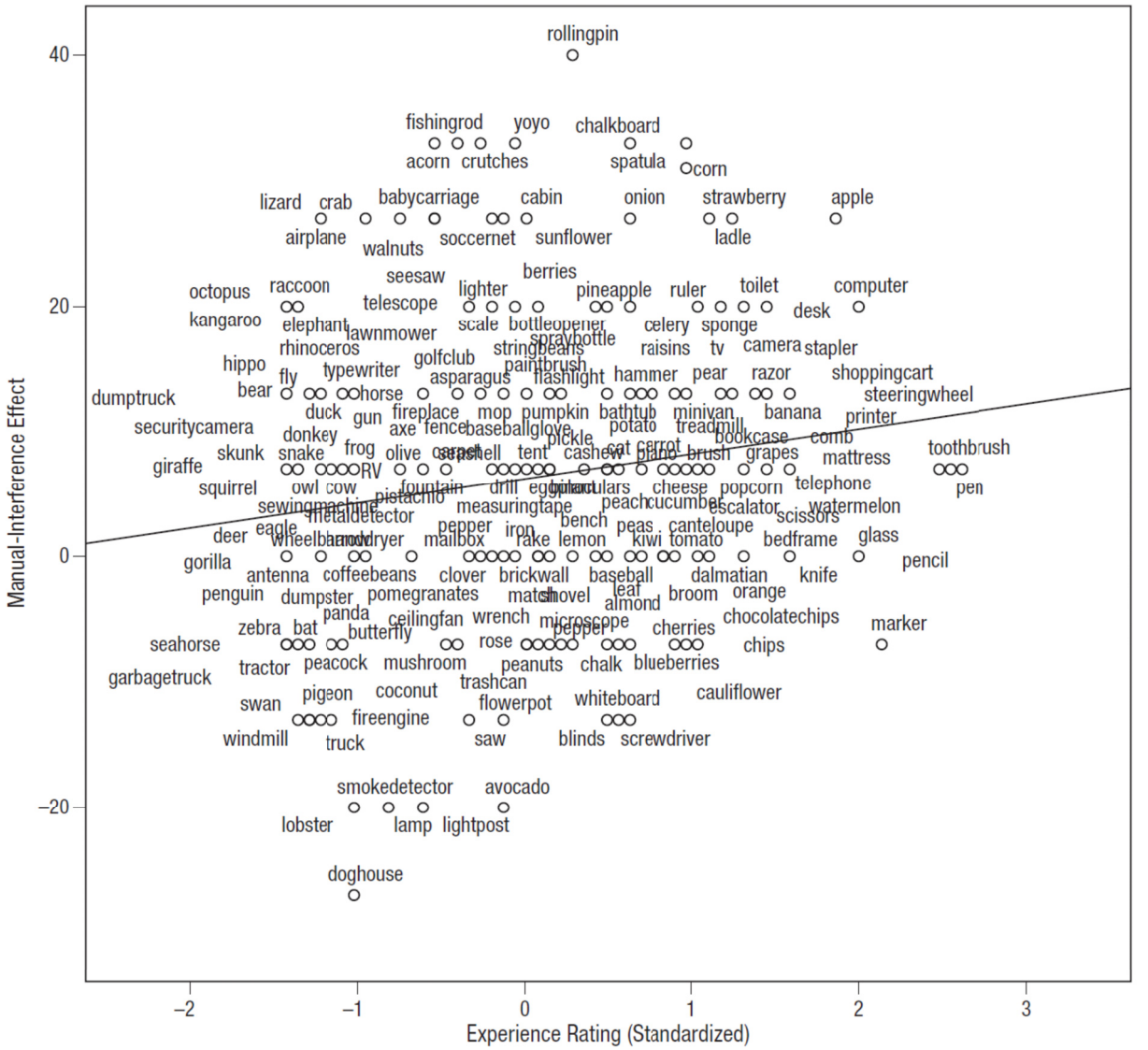
Finally, the stimulus set contained only few objects that can be reliably considered as manipulable (even less as FM), and these do not appear to have been consistently affected by the manual interference task (Figure 1). Overall, this study's results are thus largely inconclusive.

2.2.4 Lexical decisions

Some studies more specifically focused on word stimuli and investigated the role of manipulability in lexical decision tasks (LDT) that require participants to determine if a presented verbal stimulus corresponds to a real word or to a pseudoword (see Section 1.2.2). In a regression analysis on megastudy data (Balota et al., 2007), Heard et al. (2019) investigated the effects of 4 manipulability-related variables on LDT performances, namely of the ease to grasp objects, the ease to pantomime their actions (same as Guérard et al., 2015, footnote 17), the number of actions that can be performed with them, and the extent to which the body can physically interact with them. None of the variables were significantly related to latencies, whereas only the ease to grasp (SM) appeared to have a facilitatory effect on accuracies. The lack of an effect on latencies is rather difficult to interpret, but the results at least suggest that processing the words did not involve the activation of FM. Myung et al. (2006) conducted an auditory LDT coupled with verbal priming. Participants first heard a prime word followed by the target word or a pseudoword. Real word targets and primes had either similar or dissimilar FM. They found that participants responded faster when the target word was preceded by congruent primes compared to incongruent ones, which they interpreted as evidence for the activation of FM knowledge during lexical-semantic processing. However, the authors also reported that this facilitatory priming effect was not found when the analysis was performed stimulus-wise instead of participant-wise, nor when the duration of words was accounted for in the statistical models – two results that appear to have been ignored in their discussion.

To our knowledge, only two studies investigated the role of manipulability in LDT while also requiring more elaborate manual actions. Bub et al. (2008) used a derivation of their SRC paradigm (Section 2.1) in which participants had to give their responses by reaching and grasping 3-dimensional shapes that were either congruent or incongruent with the SM or the FM associated with the words' referents. In two experiments, they found that both SM and FM could facilitate LDT latencies. Their procedure nevertheless suggests that FM were elicited before SM, implying that only the former constitute a word's conceptual representation. Unfortunately, the authors did not compare their results to a condition in which participants were not required to perform a reach-and-grasp. It is thus difficult to know if the effect was due to their task's action requirements drawing attention to action-relevant features. In a rather unusual experiment, Rueschemeyer et al.

Figure 1
The manual interference task's effect on naming errors against the subjective experience ratings in Yee et al. (2013)



(2010a) hypothesised that executing a voluntary manual action during word processing would facilitate the recognition of words denoting FM objects, but not those that can only be SM. Participants performed a go/no-go LDT in which they were required to read the real words aloud. In half of the trials, they traced a circle on the desk with the index finger of their dominant hand (active condition). In the other half, their index finger was placed on a motorised rotating disk that made them passively do the same action (passive condition). Participants were more accurate in their responses to FM objects in the active condition compared to the passive one, but no difference was observed

for SM objects. In a second experiment, the oral response was changed to 'yes' and 'no' to obtain more accurate reaction times. As the rhythmicity of the active condition might have influenced the latencies, the manual task was additionally changed so that participants were either instructed to do nothing (control condition), or to simply apply constant pressure on a button (motor condition). Accuracies were generally higher in the motor condition but no interaction was observed with the type of stimulus. For latencies, they notably found that the motor condition led to faster responses for FM objects compared to SM ones, while no such difference was observed in the control condition. Thus, actions that were unrelated to those associated with the use of objects appeared to facilitate the recognition of MO names, even when they did not require planning or execution.

These results are highly curious because an irrelevant manual task during object processing is generally assumed to interfere with the motor processes involved in conceptual representations, not to facilitate them. Additionally, the second experiment's finding – if reliable – raises the question of how holding a button might affect responses in other experimental settings. Indeed, the vast majority of experiments requiring a reach-and-grasp response to a stimulus asked their participants to keep a button pressed until they made a response (e.g. Bub et al., 2008, 2021; Canits et al., 2018; Kalénine et al., 2014). If this simple action is found to have a robust effect on processing manipulable objects, then these latter findings should likely be reinterpreted and studied at more length to disentangle the experimental procedure's contribution to the effects. Interestingly, Rueschemeyer et al.'s (2010a) findings have been mostly ignored, with the study being sometimes cited in support of the opposite effect (e.g. "motor interference actually hampers object identification", p. 240, van Elk et al., 2014) or confused with another study by the same authors (Rueschemeyer et al., 2010b. E.g. Binkofski & Buxbaum, 2013; Lee et al., 2013;). Although we remain cautious in drawing strong conclusions from a single study, these results and observations reinforce our suspicion that the widely used keypress responses might also introduce unwanted confounds.

2.2.5 Semantic categorisation

Finally, several studies have asked participants to perform a semantic categorisation task which should elicit deeper conceptual processing. Most of the experiments presented below nevertheless did not clearly focus on functional manipulations despite their presumed role in conceptual representations. When not specified, MO will refer to invariably SM and FM objects. The studies in this section have been mostly inspired by Tucker and Ellis's (2001, 2004) natural/artefact categorisation experiments briefly mentioned in Section 2.1.2. As a reminder, these authors used a hand-held response apparatus that mimics precision and power grips which allowed them to manipulate the congruence between the response and the grasp that would be necessary to pick objects up. They found that participants were generally faster to categorise objects and their names when

the two grasps were compatible than when they were not (the compatibility effect) and argued that both bottom-up visual information and stored knowledge can drive the effects through the activation of motor representations. More recently, Kalénine et al. (2014) used a similar procedure to investigate whether the visual context in which the objects are shown modulates the effect. MO that required different grips to move and to use (e.g. kitchen timer) were shown in visual scenes that implied either their use or transport (e.g. on a countertop with food, or in a drawer, respectively). Participants had to decide if the objects were manmade or natural by performing a reach-and-grasp response that could be congruent or incongruent with the object's context on a 3-dimensional shape. As predicted, a significant compatibility effect between the grasp implied by the scene and the one used to respond was found, leading to the conclusion that motor information can contribute to conceptual processing but that its effects are context-dependent.

As discussed in Sections 1.3.3 and 2.1, the use of reach-and-grasp responses unfortunately does not readily allow to draw conclusions about the nature of conceptual representations from such tasks as they likely lead to task-induced effects. We have also mentioned that Tucker and Ellis's (2001, 2004) results show some inconsistent patterns despite statistically significant results, with the compatibility effect often found variably for only one of the two grasp responses. Furthermore, the authors did not conduct post-hoc analyses and sometimes did not provide information about the variance of their data, making their findings generally difficult to interpret. Kalénine et al. (2014) similarly found that reaction times were only different across contexts when participants made a pinch, but not a clench, response. Additionally, this difference was only found in a participant-wise analysis but was absent in a stimulus-wise one, and no exclusion criteria for fast and slow latencies were used – which might have yielded null results given the high variance of their data in the significant pairing. More recently, Haddad et al. (2024) used a similar paradigm (without the contextual manipulation) to investigate whether the observed effects can be reliably attributed to an activation of motor representations or if they can be explained by more abstract mappings between the stimuli and the responses (see Section 2.1). Similar to the findings described above, they found an overall compatibility effect that only emerged for one of the responses. Participants were faster to categorise large objects compared to small ones through a power grip response, while no effect was found for precision grips. Critically, however, the study also included small and large non-manipulable objects[18], and the same effect was found. Even more surprisingly, participants in another condition were asked to instead respond with incongruent gestures instead of grasps (i.e. touching the response device with the back of the whole hand or of one finger), which also yielded similar compatibility effects. As the authors argued, these results strongly suggest that such

[18]Manipulability was assessed in a pilot study by asking participants to rate the ease with which they can manipulate each object.

compatibility effects cannot be attributed to motor representations involved in object processing.

A few studies investigated the semantic categorisation of objects with keypress responses after priming them with representations of hands in congruent or incongruent postures but have also yielded inconclusive results – in addition to once again not being very informative overall due to potential task-induced action effects. Borghi et al.'s (2005, 2007) experiments required participants to categorise small and large MO as natural or artefact after seeing pictures of hands in a precision or a power grip posture. They found a significant compatibility effect, but only after participants were trained to imitate the grasps prior to the main experiment. Godard et al. (2019), on the other hand, found a similar result without a training phase. They primed MO selected based on Salmon et al.'s (2010) SM ratings (see Section 2.2.2) with either a hand with a congruent SM grip or in a neutral posture (e.g. palm down, fist) and found a compatibility effect only for manufactured objects. Presuming that these are more associated with FM than natural objects (fruits and vegetables), they concluded that the semantic categorisation of artefacts evoked their FM. However, the effect was once again found for only power grasp primes. As their stimulus selection procedure also suggests, an inspection of their stimuli additionally shows that artefacts were arguably not associated too strongly with FM (e.g. *bowl*, *coin*, *plate*, *ring*). It is thus questionable whether an activation of FM was responsible for the observed results. Vainio et al. (2008) instead primed similar MO with videos of hand grasps and asked them to perform the same categorisation task through oral and keypress responses, as well as with Tucker and Ellis's (2001, 2004) apparatus. In the two latter cases, responses to large objects were faster than to small ones when they were preceded by a power grasp video. No differences were found for objects primed with a precision grasp. Surprisingly, the precision grasp videos led to longer categorisations for small compared to large objects when participants were not engaged in a manual task (i.e. oral responses) – an effect that was completely opposite to their hypotheses and other results, but was taken as evidence for the evocation of motor representations by the objects regardless. Somewhat similar to Haddad et al. (2024), Matheson et al. (2014b) used FM artefacts and non-manipulable natural stimuli (i.e. animals), and found that a hand prime of a congruent size with the stimuli facilitated the categorisation of both artefacts *and* animals compared to a hand prime that was too small to allow interaction. It is thus unlikely that the effect was due to an activation of motor representations, at least not of FM.

To our knowledge, only three studies investigated the role of manipulability through other semantic decisions, two of which included a manual interference task. Heard et al. (2019) performed regression analyses to investigate the effect of four variables related to SM and FM (i.e. BOI, ease to grasp, ease to pantomime, number of actions) on concrete/abstract categorisation performances of words (Pexman et al., 2017). They found that all of them had a facilitatory effect on both latencies and accuracies, thus suggesting that they played a role in semantic processing. Yee

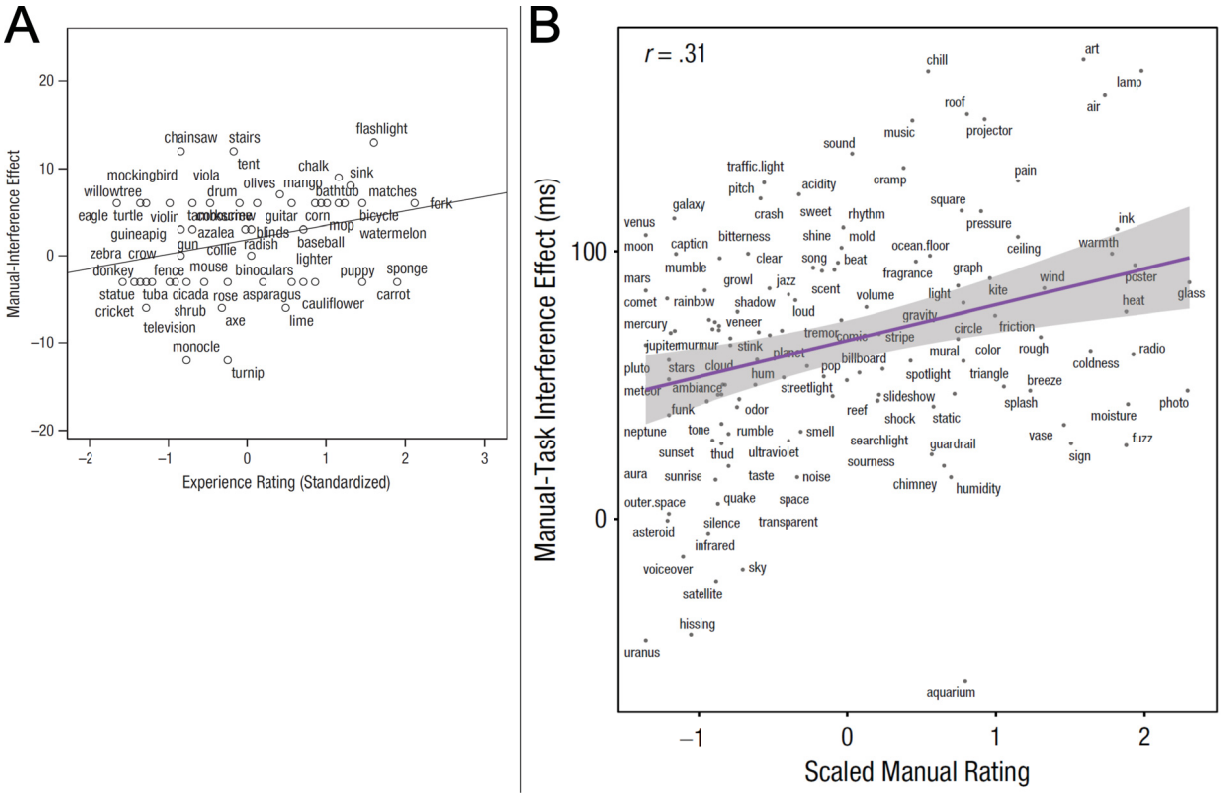
et al. (2013) similarly performed a concrete/abstract task with words (oral responses) and asked their participants to perform a concurrent manual task in some of the trials. Additionally, the participants were asked to rate each item relative to whether they have more experience looking at them or touching them. As with their naming task discussed previously (Section 2.2.3), they found no evidence that the manual task affected categorisation latencies for items with which participants had more manual experience. A small effect was nevertheless present on accuracies. Note that, in this experiment, the same participants performed the task and provided the experience ratings. However, this study's results are unfortunately once again inconclusive as the stimuli similarly included very few MO, for which no specific interference effect can be observed (Figure 2.A). Finally, Davis et al. (2020) used a similar approach but with a go/no-go "is it an animal?" categorisation of words. Participants performed the same interference task as Yee et al. (2013) during half of the trials and responded with their feet. The manual task's interference effect was assessed relative to the amount of experience participants had touching the items. This time, the interference effect on latencies (but not accuracies) was found to increase as a function of manual experience – i.e. responses were slower to items associated with more experience when performing the manual task. As with Yee et al.'s (2013) naming study, however, it was once again not the same participants who participated to the categorisation experiment that provided the experience ratings. Given also the stimuli and their respective interference effects, the claim that the results "provide strong evidence for a fundamental prediction of sensorimotor-based models [...] – conceptual representations are grounded in sensorimotor *experience*" (p. 514, emphasis in original) appears questionable.

2.2.6 Synthesis

The evidence for the involvement of either SM or FM in object processing across the different tasks presented in this section appears to be generally inconclusive. We have indeed seen that a large number of them are subject to potential task-induced action effects (i.e. hand representations, read-and-grasp actions), while others have either provided inconsistent results or are difficult to interpret due to methodological factors. A major limitation of most of these studies is notably that they have used only MO, without any non-manipulable objects. Although limited to semantic categorisation, Matheson et al.'s (2014b) and Haddad et al.'s (2024) studies are excellent examples of the danger of drawing conclusions about the involvement of motor representations specific to MO in conceptual processing without such controls. Thus, even the studies reporting robust results require further investigation before their effects can be more confidently attributed to the processing of MO.

Figure 2

The manual interference task's effect on errors (A) in Yee et al. (2013) and on latencies (B) in Davis et al. (2020) against participants' subjective manual experience ratings



2.3 Memory

The role of motor information in the representation of manipulable objects (MO) has also been investigated in memory, although by a relatively smaller number of studies compared to the other paradigms (for reviews, see also Pecher et al., 2021; Zeelenberg & Pecher, 2016). The general hypothesis in this line of research is that if processing manipulable objects activates the actions they are associated with, this motor information should be encoded in memory, increase the distinctiveness of the memory trace and provide additional cues during retrieval. In what follows, we present the available evidence for the role of motor information in the memory for MO across two broad paradigms, namely short-term (or working) memory, and long-term memory. In the current context, MO will be used to refer to objects that are functionally manipulable, unless specified otherwise.

2.3.1 Short-term memory

Pecher, Zeelenberg and their collaborators conducted a series of interference studies to determine if the availability of the motor system modulates memory performance for MO (but not for non-manipulable objects – NMO). Pecher (2013b) used a task in which participants saw two images of an object separated by a 5000 ms interval. On the second presentation, they had to decide if the picture was the same or a mirror image of the first stimulus by pressing pedals with their feet. In one condition they additionally had to continually make a fist with both of their hands, extend their fingers one by one, and then repeat until the end of the trial (finger flexion task). Contrary to the prediction that motor information is causally involved in maintaining MO in memory, they found that motor interference trials did not affect performance more for MO than NMO. Follow-up experiments similarly showed that the manual task did not interfere with memory for MO neither when the experiment required more specific identification of the object, when it was more difficult, nor when participants could not rely on bottom-up visual information.

Pecher et al. (2013) found similar results with the same interference condition in an N-back task in which participants were presented with a sequence of object pictures and had to respond by foot when one of them was the same as a previous one, variably 1 to 4 trials back. Finally, Quak et al. (2014) repeated the N-back task by further distinguishing MO into those that can be grasped with a precision grip and those with a power grip. In the motor interference trials, participants had to rhythmically squeeze either a small foam cylinder with a precision grip or a larger foam cuboid with a power grip. The authors found no specific interference effects of the different manual tasks on the performance for objects with corresponding grips. Summarising 11 experiments conducted by their team through a small meta-analysis, Zeelenberg and Pecher (2016) concluded that "in recognition-like short-term memory tasks, there is no evidence that motor simulations support memory performance" (p. 20).

Contrasting the above results, Guérard, Lagacé and collaborators consistently found motor effects in the memory for MO through different tasks. Guérard and Lagacé (2014) noted that the experiments conducted by Pecher (2013b) and Pecher et al. (2013) used only 'pure lists', i.e. MO and NMO were never presented together[19]. They argued that if objects are manipulated (i.e. grasped or used) in similar ways within a list, motor information might be involved without being distinctive – and useful – enough to perform the task (on distinctiveness, see e.g., Nairne, 2002). To test this hypothesis, they combined an isolation paradigm with a motor interference paradigm. The

[19]One exception is the second experiment by Quak et al. (2014) in which the experimental lists contained NMO and MO that can either be grasped with a precision or a power grip. The authors did not find a motor interference effect on the performance for MO as with all other experiments. However, this result does not directly extend to their previous experiments as the focus of this study was specifically on two types of structural manipulations, whereas functional actions might be more relevant to memory.

isolation effect (or Von Restorff effect, Von Restorff, 1933) is a tendency for superior recall for an item when it possesses a characteristic that differentiates it from the rest of the list. For instance, if the list contains 5 blue shapes and 1 green shape, the green shape would be better recalled because of its distinguishing feature (note that if all shapes were blue, colour might be processed but would not be useful for recollection). The authors showed two pure lists of MO and NMO respectively, as well as a list with an isolated MO and a list with an isolated NMO. Participants had to recall the items of a given list in the correct order (serial recall). Isolates were found to have higher correct recall rates compared to the items in the same position of the pure lists. Critically, this advantage disappeared in a subsequent experiment in which participants had to do a concurrent finger flexion task (i.e. motor interference), thus suggesting that it is motor information associated with the objects that was responsible for the effect.

Downing-Doucet and Guérard (2014) and Lagacé and Guérard (2015) used a different protocol in which short videos of a hand performing several types of grasps to pick up and move objects (structural) were followed by photographs of MO. Downing-Doucet and Guérard (2014) used grasps that were always congruent with the objects, but varied their type – and thus their distinctiveness. In one condition, the same grasp was presented before each object, while 4 different grasps were shown in a second condition. Serial recall rates were higher when the grasps were different and, most importantly, the difference was abolished when participants performed a finger flexion task. Lagacé and Guérard (2015) showed different types of grasps that could either be congruent or incongruent with the object and asked their participants to simply watch the videos (control condition) or to perform the same grasp on a 3-dimensional object. They found an effect of grasp congruency on serial recall performances, but only when participants executed the grasps. A second experiment replaced the videos with pictures of MO that could be used in a similar or dissimilar way to the target objects. This time, participants had to pantomime the use of the first object or do nothing. They similarly found an effect of congruence, but only when participants made an overt action. Overall, the studies presented by the two teams lead to conflicting interpretations, but also contain a number of limitations that make it difficult to draw reliable conclusions. As mentioned, the experiments by Pecher and collaborators have predominantly used pure lists which might undermine the distinctiveness of the items. A quick glance at their stimuli nevertheless reveals that the MO in their lists were generally associated with both different functional and structural manipulations. It thus appears unlikely that motor information was entirely irrelevant as cues to discriminate the objects. On the other hand, Downing-Doucet and Guérard (2014) and Lagacé and Guérard (2015) used protocols that explicitly referred to manual actions (hand postures and overt action execution) which might have primed attention to the stimuli's action-relevant features. The conclusions that can be drawn about the inherent role of motor information in the

representation of MO from these experiments is thus limited. However, Guérard and Lagacé's (2014) isolation experiment appears to be a more robust result that is difficult to reconcile with the findings of Pecher and collaborators. As Zeelenberg and Pecher (2016) have argued, it is possible that the discrepancy is due to the type of task used during recollection (recognition vs. recall) but how this might have affected the results is not entirely clear. An alternative explanation could be that Guérard and Lagacé (2014) asked their participants to name the objects during the learning phase. As we have seen in Section 2.2.3, however, the evidence for the role of either SM or FM in naming is not well established.

2.3.2 Long-term memory

Evidence for the role of motor information in long-term memory for MO is similarly mixed. Canits, Pecher, and Zeelenberg (2018) revisited the grasping compatibility paradigm used by Tucker and Ellis (2004. Section 2.1.2) and added a surprise free recall phase[20]. Participants performed a natural/artefact categorisation by reaching and grasping one of two 3-dimensional shapes that could either be congruent or incongruent with how a given object would be picked up (structural). After the task, they were unexpectedly asked to recall as many objects as possible (free recall). The authors hypothesised that incongruent grasps would produce motor interference, thus resulting in an impoverished memory trace and reduced performance compared to congruent trials. While the compatibility effect was replicated for reaction times, no effect was found on recall performance, whether using object pictures, their names, or a recognition task instead of free recall. The authors concluded that the results on reaction times are consistent with the view that objects evoke their associated actions, but that the lack of an effect on memory performance suggests that these were not encoded in memory. However, this study also presents some important limitations. First, participants performed rather arbitrary reach-and-grasp actions on simple cylinders that were likely not congruent with how the objects would have been grasped for their use. The response modality could have thus biased attention towards structural manipulations and have interfered with the activation of functional ones. In such a case, the association of only two actions with the 80 stimuli used in their experiments would have likely not represented distinctive enough cues during recollection. Second, participants were engaged in a manual task throughout the experiment by keeping a button pressed until they were ready to reach and grasp one of the cylinders. It is thus also possible that reaction times were affected only by an effect on the preparatory phases of the reach-and-grasp action or simply by pressing the button (Section 2.2.4), and that the objects in themselves did not evoke any actions due to the interference.

[20]Note that the MO used by this study were structurally, but not always functionally, manipulable, especially in the case of natural objects.

Other studies circumvented the difficulties that arise from manual responses. Dutriaux and Gyselinck (2016; Dutriaux et al., 2019a) presented participants with several stimulus lists composed of MO and NMO, and manipulated their posture while they memorised the items. For half of the lists, participants were instructed to learn them while their hands were at rest on the desk and they were told that they could move them if necessary (control posture). The other half required them to hold their hands behind their back (interference posture). After each list, participants performed a distractor task followed by oral free recall with their hands at rest. With both object pictures and their names, an effect of the interfering posture was observed only for MO. This result was replicated by Onishi and Makioka (2020) who further showed that placing either a transparent or an opaque board on participants' hands to constrain them on the desk yields similar interference results. In a subsequent study with words, Dutriaux and Gyselinck (2021) also replicated the effect by manipulating the posture during recall instead of the learning phase. This result is particularly interesting as it shows that interfering with access to motor information during retrieval also reduces performance for MO, thus strongly suggesting that it was part of the memory trace. However, a similar and better powered experiment by Pecher et al. (2021) failed to find the same effect. This study mainly differed from Dutriaux and Gyselinck's (2021) procedure in that it used an active interference task and object pictures instead of words. Participants sequentially touched their thumb to each of their other fingers with both hands during the free recall task. The authors found no evidence that the concurrent manual task affected memory for MO more than NMO. It is difficult to determine whether the different interference tasks, the use of object pictures instead of words, or another factor were responsible for the contradictory findings. We can nevertheless note that if motor interference selectively reduces memory for MO pictures during learning (Dutriaux & Gyselinck, 2016; Onishi & Makioka, 2020), then one would expect to find the same effect by interfering with the motor system at recall as well. Regarding the interference condition itself, it is possible that Pecher et al.'s (2021) task drew attention away from the stimuli. Indeed, the authors found that the concurrent action resulted in reduced performance for both MO and NMO, whereas the other studies only reported an effect on MO. Given that no floor effect appears to have been involved, however, a stronger interference for MO would have still been expected.

A few concerns also cast some doubt on the generalisability of the significant interference effects reported above. First, Dutriaux and Gyselinck (2016) performed their published experiments after conducting several pilot studies to find a procedure that reliably produced an effect (personal communication[21]). This does not undermine their findings in any way as there are no

[21] We obtained this information from a member of Léo Dutriaux's thesis committee at the *LabEx CORTEX: The dynamic and flexible nature of memories* conference in Lyon, France (2019, June), while discussing our failure to replicate the effect with a slightly different experimental design.

apparent methodological issues, and their results have been replicated on several occasions. It nevertheless suggests that the interference effect might be particularly difficult to observe and opens the questions of which procedures did not work and why. Regarding Onishi and Makioka's (2020) study, their method suggests that participants in the control condition had to keep their hands on the desk in a stricter manner than in Dutriaux and Gyselinck (2016). Although their hands were not physically constrained, the instructions indeed implied that they could not move them, which could be expected to affect the availability of the motor system as well. In line with this interpretation, we have previously found preliminary evidence across two memory experiments that simply asking participants to keep their hands in a fixed position (on their lap, in our case) affects memory performance for MO (Paisios et al., 2019). Given the results presented by the authors, it is also doubtful that a significant effect would have been found if any of the interference conditions were compared with the control condition separately. Their analyses only report a difference when the control condition is compared with the combined results of their three interference tasks. Overall, the evidence from interference studies for a role of motor information in the long-term memory of MO is thus not entirely convincing.

To our knowledge, only two studies investigated how different contexts affect memory for MO. Dutriaux et al. (2019b) used the same postural interference paradigm presented above. Instead of MO and NMO, however, participants were only shown MO in two different verbal contexts. Objects could be either associated with an attentional verb (e.g. *to examine a hammer*) or with an action verb (e.g. *to hold a hammer*), and participants memorised the names of the objects. In line with a context-dependent recruitment of motor information, they found that the postural interference drastically affected free recall rates of MO presented in an action context, whereas no difference was observed for those in an attentional context. As acknowledged by the authors, it is nevertheless not directly possible to draw the conclusion that motor information associated with MO was responsible for the effect. The experiment indeed lacked a control condition with NMO, and it is possible that the postural interference affected participants' ability to process the action contexts more generally.

Madan and Singhal (2012) predicted that the names of objects associated with functional manipulations (FM) should be retained better than those that can only be structurally manipulated (SM). Their experiment further involved three conditions to investigate how manipulability's effect changes relative to how deeply the stimuli are processed. In the 'shallow' processing condition, participants had to determine if the words had an even or odd number of letters. The 'intermediate' condition required them to judge if the objects could be functionally manipulated with their hands. Finally, participants in the 'deep' condition were asked if they had seen the objects during the last three days. A surprise free recall task was then administered. FM objects had higher recall rates

than SM ones in the 'shallow' and 'deep' conditions. Surprisingly, less FM than SM objects were recalled in the 'intermediate' condition involving manipulability judgements. The authors argued that an automatic activation of motor representations was responsible for the first two effects, but that the conscious processing of manipulability likely led to allocate more attentional resources to SM objects because participants "could not as easily imagine interacting with them in the first place" (p. 1569). There are unfortunately several issues that do not readily point to this conclusion. First, it is not evident why it would be more difficult to decide that an object cannot be functionally used than to decide that it can, and participants took on average significantly longer to judge items as FM than not FM (161 ms). Most importantly, the stimulus used lists did not, in fact, represent FM and SM objects (personal communication[22], March 14, 2019). The SM list predominantly contained objects that do not afford structural actions (e.g. *barn*, *moon*, *tractor*, *smoke*), as well as others that could be ambiguous regarding their functional use (e.g. *apple*, *basket*, *speaker*, *vase*). The FM list similarly contained several potentially ambiguous items (e.g. *balloon*, *doll*, *fridge*, *watch*). Additionally, the analyses were not performed on these initial lists, rather on new lists determined from a median split of participants' judgements in the task. These are reported to have had a correlation of $r = .71$ with the initial categorisation (i.e. 50% explained variance), which effectively indicates that a lot of items were not strongly agreed on being FM or not FM. As no further data is available, it is not possible to determine how ambiguous items were categorised in the analyses. It can nevertheless be suspected that they had a strong impact on manipulability judgements and on memory performance. Note that it is also not possible to conclude that motor representations drove the effects in the 'shallow' and 'deep' conditions as the 'SM' list was composed of a very heterogeneous set of items that did not differ from FM ones strictly regarding their manipulability. Overall, the question of how decision requirements affect memory for MO thus remains largely open.

2.3.3 Synthesis

Compared to the previous section, the advantage of memory studies is that they have generally used more robust paradigms, notably with a comparison of MO to NMO and the use of manual interference tasks (see also Ostarek & Bottini, 2021). Additionally, task-induced motor effects were arguably less likely as most experiments did not use procedures that explicitly drew attention to manual actions, nor required manual responses. As we have seen, however, this section similarly provides a mixed picture about whether manipulation knowledge contributes to object processing

[22]Email exchange with Christopher Madan to request the stimuli used in their experiment.

and memory. It should nevertheless be reminded that only a relatively small number of studies and teams have investigated the question through memory paradigms.

2.4 The Body-Object Interaction effect

Body-Object Interaction (BOI) is a psycholinguistic variable inspired by embodied cognition and was introduced by Siakaluk et al. (2008a) to assess the general sensory-motor richness of our experiences with entities. As its name suggests, participants are asked to rate "the ease or difficulty with which a human body can physically interact with the word's referent" (p. 437), typically on a 7-point ordinal scale with higher values representing easier interactions. The emphasis on physical interactions makes this dimension particularly sensitive to motor information associated with objects. Heard et al. (2019) showed that a large portion of the variance in BOI ratings can be explained by various manipulability-related ratings such as the ease to grasp objects, how easily their use can be pantomimed and the number of actions that can be performed with them (see Section 2.2 and Chapter 4). BOI thus appears to closely reflect object manipulability – albeit on a rather coarse-grained level. The variable's effect has been investigated across a large variety of tasks and populations since its inception, mostly thanks to the wealth of rating datasets made available (Alonso et al., 2018; Bennett et al., 2011; Bonin et al., 2013; Lalancette et al., 2024; Muraki et al., 2022; Paisios et al., 2023; Pexman et al., 2019; Tillotson et al., 2008; Villani et al., 2019). A large portion of studies have shown that words with high BOI ratings (e.g. *hammer*) are processed faster and more accurately than low BOI words (e.g. *cloud)* . This facilitatory *BOI effect* has been generally cited in support of a sensory-motor grounding of knowledge and interpreted as showing that richer semantic representations facilitate word recognition (Section 1.2.2). However, there have also been some conflicting results that have been difficult to reconcile with the proposed theoretical accounts.

2.4.1 Lexical tasks

One of the main approaches to investigating BOI's effect in word processing has been through the lexical decision task (LDT, Section 1.2.1). Several studies have reported a facilitatory effect of the variable on LDT (including auditory LDT) latencies in both adults and children (Bennett et al., 2011; Hansen et al., 2012; Siakaluk et al., 2008a; Taikh et al., 2015; Tillotson et al., 2008; Xu & Liu, 2024a, 2024b; Yap et al., 2012; de Zubicaray et al., 2023). In contrast, however, a comparable number of studies have found null results (Hargreaves & Pexman, 2014; Heard et al., 2019; Juhasz et al., 2011; Lund et al., 2019; Muraki & Pexman, 2021; Pexman et al., 2019;

Taikh et al., 2015; de Zubicaray et al., 2023) and a few have even reported an inhibitory effect, i.e. high-BOI words recognised slower than low-BOI ones (Alonso et al., 2018; Lalancette et al., 2024). In a short review, Connell and Lynott (2016) argued that one reason for these discrepancies might be a confound between BOI and the age of acquisition (AoA), as physical objects which can be interacted with are likely acquired earlier in life. AoA has been consistently shown to affect word recognition performances (e.g. Ferrand et al., 2011; Kuperman et al., 2012) but has not been controlled in several experiments on BOI. Unfortunately, studies that have controlled this variable do not provide a clearer picture and display the same pattern of contradictory results (e.g. Alonso et al., 2018; Bennett et al., 2011; Heard et al., 2019; Juhasz et al., 2011).

A highly interesting and curious finding was reported by Pexman et al. (2019) in the largest analysis to date with 3591 nouns. After controlling for AoA and several other critical variables, the authors found a quadratic BOI effect on LDT performances with shorter latencies (and higher accuracies) for midscale words compared to those at the two ends of the scale, but no clear differences in performance between items at the extremes. Given its breadth, this study highly suggests that low and high BOI words do not differ markedly in the ease with which they are recognised. It is nevertheless important to stress how counterintuitive the quadratic effect is; there is no a priori theoretical reason why objects that are moderately easy to interact with should be processed faster and more accurately than those with which it is very difficult or easy to do so. We will explore a potential explanation for this finding in Chapter 3.

Although words with richer semantic representations are expected to facilitate word recognition, it is possible that the LDT is simply not sensitive enough to semantic effects for BOI to play a role (Section 1.2.2). A few studies have investigated BOI's influence in word pronunciation which similarly requires little semantic processing and have also yielded conflicting results. Bennett et al. (2011) and Yap et al. (2012) have found a facilitatory effect on pronunciation latencies, whereas Alonso et al. (2018) and Wellsby and Pexman (2014) reported null findings. Larger-scale studies would be required for more robust conclusions about this task, but the similarity of the findings to those with the LDT suggests that the BOI effect is indeed highly unstable in semantically shallow tasks. However, this interpretation does not readily explain the divergent findings. In the same vein as Connell and Lynott's (2015) suggestion for AoA, one possibility is that they arise due to confounding factors. Indeed, studies sometimes differ considerably in the variables included in their analyses or used for their stimulus selection criteria, and whether BOI's effect changes with respect to them has not been systematically studied. As an example, Lalancette et al. (2024) showed that

including imageability in the analysis of LDT latencies results in a reversal of BOI's effect[23] (inhibitory with imageability, facilitatory without), which is additionally contingent on the range of values included for each variable. Without a better understanding of such interactions and of how they affect the statistical models, it is exceedingly difficult to draw reliable conclusions about any variable's effect.

2.4.2 Semantic tasks

As a semantic variable, BOI's effect should be more evident when words' meanings are more directly accessed and in contexts in which the dimension is useful. There has been a wealth of experiments using semantic tasks – and particularly semantic classification (SCT) – that overall point to a robust BOI effect in some cases, but that also suggest that its effect varies depending on a given situation's decision requirements. The most widely studied case has been when participants are asked to classify words as concrete or abstract. With the exception of Hargreaves and Pexman (2014) who used a difficult protocol with imposed short speeded response times (75, 100, 200 and 400ms), all studies analysing performances in this task have reported a facilitatory BOI effect (Bennett et al., 2011; Heard et al., 2019; Pexman et al., 2019; Taikh et al., 2015; Yap et al., 2012; de Zubicaray et al., 2023). Newcombe et al. (2012) split this task by asking two groups of participants to respectively respond to either concrete or abstract entities only (go/no-go paradigm). Interestingly, they found that words with higher BOI values were associated with faster and more accurate responses when participants classified objects as concrete, but that the reverse pattern emerged for abstract responses (i.e. slower and less accurate responses to higher BOI words). In support of the sensory-motor grounding of knowledge, the authors argued that motor information as captured by BOI is diagnostic of concrete entities, which facilitates their processing in the concrete task. In the abstract condition, however, this information would not be congruent with the decision and its activation in words with higher BOI values was thus suggested to interfere with categorisations.

Similar results have been found in SCTs focusing on imageability. Several studies used a go/no-go paradigm in which participants made a response when a word referred to something that is easily imageable, and reported a facilitatory BOI effect on latencies (Duffels, 2022; Hansen et al., 2012; Hargreaves et al., 2012; Wellsby et al., 2011. For similar results with a two-choice paradigm, see Siakaluk et al., 2008b). Wellsby et al.'s (2011) study is particularly interesting because it tested a crucial hypothesis that, to our knowledge, has not been directly investigated otherwise.

[23] Although not presented in the manuscript, our own analyses with the BOI ratings introduced in Chapter 3 have yielded similar results. We compared two models which included a large number of control variables, and differed with respect to the inclusion of either imageability or concreteness. When imageability was controlled, BOI had an inhibitory effect on LDT latencies, whereas its effect disappeared when concreteness was included instead.

The authors noted that, as experiments predominantly use keypress responses, BOI's effect could be explained by task-induced motor priming instead of the activation of motor information resulting from semantic processing. To test this hypothesis, they performed three experiments using the same go/no-go imageability task while varying how participants responded. In one condition, participants used a typical keypress response. The other two conditions involved verbal responses, with either the pronunciation of the easily imageable words, or simply saying "yes" to more closely match the first condition. In all cases, responses were faster to high (vs. low) BOI words with no interactions between conditions, thus strongly suggesting that the effect is due to the involvement of motor knowledge instead of resulting from motor priming. Note, however, that this interpretation poses an important challenge. If BOI primarily captures manual motor information which, as the study suggests, is involved during semantic processing, then why does it not interact with keypress responses? The authors pertinently predicted that such an interaction could be facilitatory due to priming. Conversely, it could also be argued that the involvement of the hands in the task can interfere with the simulation of manual motor information. In both cases, however, some interaction would be expected between verbal and manual conditions because keypress responses and semantic motor simulations should draw on the same resources. The involvement of motor information in this task is thus not as evident as it appears at first glance.

BOI's effect has also been investigated in a number of other semantic tasks, the results of which are generally taken to reflect the flexibility of semantic processing and its modulation by context. For instance, Tousignant and Pexman (2012. See also Muraki et al., 2023b) conducted four semantic classification tasks with the same set of stimuli, but different instructions. Participants were instructed to classify words as: (1) action/non-action, (2) entity/non-entity, (3) entity/action, or (4) action/entity. The critical manipulation in this study was whether participants were explicitly informed of the stimuli's categories. In condition (1) in particular, no indication was given that object-related knowledge could be relevant for the task. The results revealed a processing advantage for high-BOI words on response latencies (although not consistently on accuracies), but only when the task required responses to entities explicitly (i.e. conditions 2–4). Additionally, the difference in latencies between low- and high-BOI words was largest when the only considered category were entities (condition 2). The authors argued that the decision requirements of the different tasks modulated participants' expectations about the information that would be relevant to solve them. Thus, BOI would not be particularly diagnostic when actions have to be distinguished from non-actions, but more relevant when the task draws attention to whether something is an entity.

The previous result strongly suggests that sensory-motor information is drawn upon depending on its expected relevance in a given context. However, it is quite difficult to determine in which contexts motor information (and by extension BOI) can be expected to play a role. For instance,

BOI appears to facilitate the discrimination of nouns from other syntactic categories (Muraki et al., 2023c; Muraki & Pexman, 2021), but not the classification of stimuli as living or non-living entities (López Zunini, 2016; Taikh et al., 2015). More surprisingly, BOI does not seem to affect decisions about whether an object is possible to touch or not, even though the task directly taps into motor knowledge (Al-Azary et al., 2022). There have also been some conflicting results that add to the complexity of determining BOI's role in semantic processing. Two such cases come from a few studies using picture naming and memory paradigms. In the former, Bennett et al. (2011) found that BOI facilitates the naming of pictures. However, subsequent studies have failed to replicate this effect (Bonin et al., 2013; Taikh et al., 2015; de Zubicaray et al., 2023). To our knowledge, only two studies have investigated BOI's effect on word memory. As we have mentioned in Section 2.3, motor information should be associated with richer and more distinctive semantic representations, thus having a facilitatory effect in memory. Marre et al. (2024) recently analysed free recall rates across four experimental conditions and found no effect of BOI when participants were asked to mentally repeat the words, visualise their referents, or to imagine performing an action with them. The variable only had a facilitatory effect in a 'situated motor imagery' condition in which the imagined actions were performed in a richer and social context. On the other hand, Lau et al. (2017) conducted a memory megastudy and found that BOI had no effect on recognition performances. Surprisingly, higher BOI words were also found to be associated with lower free recall rates.

2.4.3 Synthesis

Overall, the general evidence suggests that BOI does not play a role in tasks requiring no direct processing of word meaning, and that it has a facilitatory effect in a select number of semantic tasks. The presented findings for the latter type of tasks also strongly point to the task-dependent nature of BOI and have been largely interpreted as such. As we have pointed out, however, it is not entirely clear when BOI can be expected to be relevant, or even if it is truly motor information that is responsible for the observed effects (see also Section 1.2.4).

2.5 Discussion and thesis outline

The literature discussed in this chapter paints a rather perplexing picture about the role of motor processes in the conceptual representation of manipulable objects. On the one hand, there is little consensus overall on whether structural and/or functional actions play a role in object processing, under what conditions they might do so, or even the most appropriate ways to experimentally investigate these questions. On the other hand, a detailed look into this literature reveals

important limitations that make it extremely difficult – if not impossible – to determine what both significant and null findings truly reflect. One would hope that further experiments will eventually shed some light on these results and bring us new insights about conceptual processing and its dynamics. However, the last 25-odd years of empirical inquiry on the topic have produced more controversial and confusing findings overall than a cohesive understanding – and there is little reason to expect this trend to change anytime soon.

Our fundamental thesis in the present work is that the discrepancies partly stem from the fact that experiments are predominantly theory-driven, while their methodological tools and procedures often take a back seat and are rarely studied themselves. For instance, it is somewhat paradoxical that the effects of different manual responses (e.g. keypresses) on object processing have not been more systematically investigated, in a literature whose core hypothesis centres on the role of manual motor representations. Another example to which we have alluded throughout our review is the general lack of consensus and discussion regarding how to operationally define *manipulability*, i.e. the key feature of the objects under study. Going further, the assessment of manipulability and other stimulus characteristics through Likert-type scales has been mostly taken for granted without much inquiry, despite being pervasive across virtually all experimental fields of cognitive science since the dawn of the discipline (e.g. Osgood, 1952). How can we expect to obtain coherent and theoretically informative results if the methods that we use to acquire them are neither agreed on nor understood?

These observations are not intended as targeted criticisms, rather to highlight a general lack of critical research into our methods; they have simply not been a focal point in our discipline's scientific culture. As with any nascent science, theoretical frameworks, promising lines of inquiry, and experimental tasks unavoidably need to be established first. After all, studying methodology in a vacuum makes little sense. The main issue, however, is that our theories are severely underspecified when it comes to guiding specific experimental design choices and for making precise predictions about how they might affect the results. In turn, achieving a finer-grained theoretical understanding necessarily relies on empirical studies, and it is at this point that we now find ourselves stuck. Without clearly established methodological guidelines and standards, researchers must rely on sometimes arbitrary decisions and procedures. This inevitably leads to extensive flexibility (or 'researcher degrees of freedom', Simmons et al., 2011; Wicherts et al., 2016) in conducting experiments, which can produce findings that are difficult to interpret and compare, or that are outright unreliable. We believe that this poses a serious risk of hindering theoretical progress and that it can lead to self-perpetuating debates on the implications of empirical findings – as has mostly been the case in the embodied cognition literature (see Section 1.1.4 and Zwaan, 2014, 2021). Given all these elements, it appears crucial that we dedicate more attention to our methods, not only to

provide a solid foundation for future research but also to shed light on past findings.

The research presented in the following chapters traces our exploration of some of the points mentioned above and can be summarised into two seemingly simple questions: what do Likert-type scale ratings really tell us about the dimension being assessed? And, how do we operationally define manipulability? In Chapter 3, we present a case study based on body-object interaction ratings to introduce a problem that arises from computing averages on ordinal scale judgements (Pollock, 2018), and to lay out some of its practical implications. We further complement our discussion through a detailed look into what standard summary statistics (i.e. means and standard deviations) represent on Liktert-type scales. Armed with new insights, we turn in Chapter 4 to an in-depth review and analysis of the extant manipulability ratings and attempt to assess their validity. Examining a large number of studies with different methodologies in this chapter additionally allows us to highlight broader methodological considerations for rating procedures that go beyond manipulability. In Chapter 5, we finally present a new multidimensional set of manipulability ratings collected in light of our review, which allows us a direct comparison of the results obtained through different instructions and to further discuss their validity.

Chapter 3

The elephant in the middle of subjective rating scales [24]

[24] This chapter is based on a peer-reviewed article published in *Collabra:Psychology* (Paisios et al., 2023). The introduction has been slightly modified and includes an additional example in Section 3.2. The analysis of the available literature was also updated to include some of the studies presented in Chapter 2 (Section 2.4) but not included in the published version. We have mostly kept the discussion as it appeared in the article. A short conclusion has nevertheless been added (Section 3.9) to extend our argument and to briefly address the implications of our findings for the studies investigating the body-object interaction effect discussed in the previous chapter.

A large portion of the experiments discussed in the previous chapter – and in experimental cognitive science in general – rely on normed materials. These are crucial for a study's validity as they allow to both control the stimuli for known confounds and to separate them along dimensions of theoretical interest. Despite their importance, however, such variables have been subject to surprisingly little methodological inquiry. Norming studies typically ask participants to rate a list of items along a theoretical dimension through a Likert-type scale. The average of all participants' responses for a given item is then computed and considered to represent its position on the variable's continuum. Pollock (2018) recently pointed out that this practice introduces an almost entirely overlooked confound: if participants disagree in their judgements for an item (e.g. some rate it on the low end of the scale, while others on the high end), then the average rating will inevitably tend towards the middle of the scale but will not be a reliable reflection of the underlying responses (see also Marcus-Roberts & Roberts, 1987). As obvious as this *midscale disagreement problem* may seem, we will see that it has been largely ignored and that it can undermine the results of a substantial number of studies.

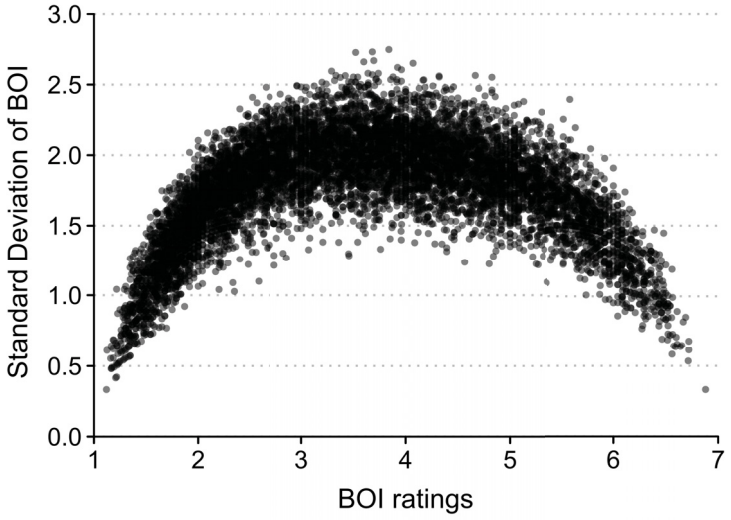
The aim of the current chapter is to unravel some of this problem's consequences through a concrete case study of the body-object interaction (BOI, Siakaluk et al., 2008a) literature presented in the previous chapter (Section 2.4). We start by using available BOI norming datasets to outline three major implications stemming from the midscale disagreement problem. We then explore the extent to which they affect the experiments on the variable's role in word processing. Finally, we present item-level descriptive and exploratory analyses based on a new set of BOI ratings for 1019 French words, which allows us to take a more detailed look into the disagreement problem.

3.1 The midscale disagreement problem

As mentioned, the midscale disagreement problem arises from averaging disparate values drawn from a bounded scale, the result of which naturally falls towards the middle of the scale. It is best illustrated by plotting the standard deviation (SD) of items against their corresponding mean ratings. SDs capture the average spread of responses around the mean and can thus be taken as a rough measure of interrater disagreement. For BOI, as for most other subjective variables examined by Pollock (2018. See also Brainerd et al., 2021), SDs display a concave relationship with the average ratings (Figure 3). Words close to the ends of the scale tend to have small SDs, indicating that raters mostly agreed about their BOI judgements. Those towards the middle of the scale, however, generally present high SDs. Such a pattern is expected to some extent as midscale response options are often not precisely defined and, more generally, because of the scale's bounded nature. The amount of observed deviation in the middle of the scale is nevertheless too high to be

solely due to these reasons and can only be explained by a significant disagreement in the ratings. For reference, a completely uniform response distribution on a [1, 7] scale yields an average rating of 4 and an SD of approximately 2. This suggests that the average rating of a large portion of midscale words does not reflect a consensus among respondents, rather that it is a methodological artefact. As a result, the ratings of such words do not fall on the variable's continuum and cannot be interpreted as representing their position on the scale.

Figure 3
Standard deviations of the BOI ratings as a function of their means for 9351 words collected by Pexman et al. (2019)

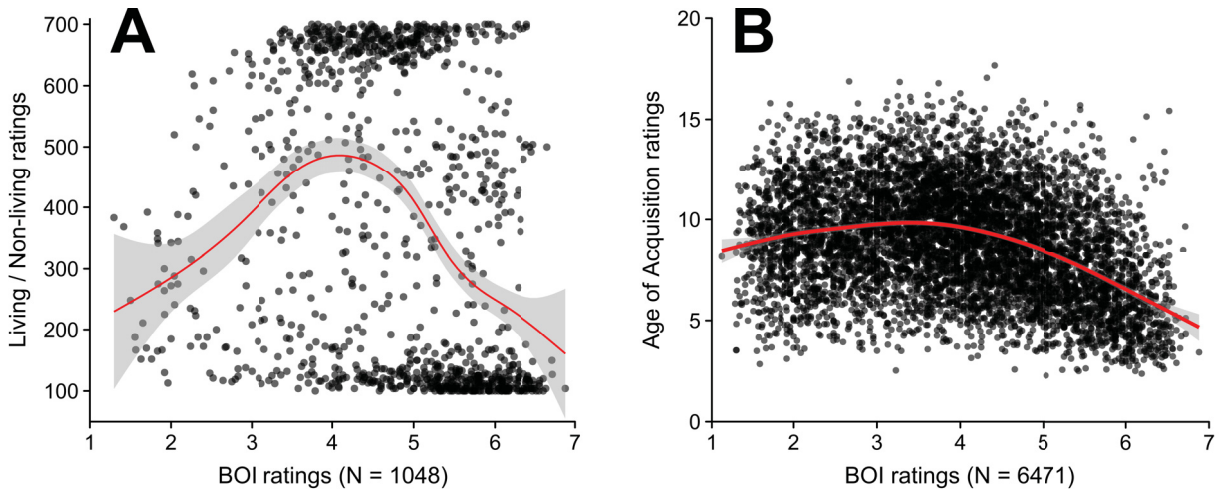


A direct consequence of the above point is that any differences in processing found for midscale items cannot be reliably attributed to the variable of interest. A thorough investigation of what generates disagreement in BOI ratings is beyond the scope of the current work. However, some simple examples show that the middle of the scale can be expected to display an independent effect driven by confound variables. The most straightforward cause is semantic ambiguity which can lead raters to interpret the same words differently and is known to affect word processing performances (Eddington & Tokowicz, 2015; Haro & Ferré, 2018. See also Brainerd et al., 2021). A quick inspection of the living/non-living ratings provided by VanArsdall and Blunt (2022) also reveals that animate entities tend to have midscale BOI ratings (Figure 4.A). Animacy's influence has been primarily investigated in memory tasks but has been recently shown to also affect word processing (Bonin et al., 2019). Following the concerns raised by Connell and Lynott (2015) about a confound between the age of acquisition (AoA) and BOI (Section 2.4), Figure 4.B shows that words with low and midscale BOI ratings display considerable variation in the age at which they are learned with a slight increase towards the middle of the scale, whereas those at the high end of

the scale are generally acquired at a younger age. If not controlled for, a combination of several such variables could lead to differences in processing performances for midscale words – without it being an effect of BOI in itself.

Figure 4

Living/Non-living (A; VanArsdall & Blunt, 2022) and age of acquisition (B; Kuperman et al., 2012) ratings against BOI ratings (Pexman et al., 2019)



Note. For Living/Non-living ratings (A), larger values correspond to living entities. The red line represents the fit from a generalised additive model and the ribbon the 95% confidence interval on the fit.

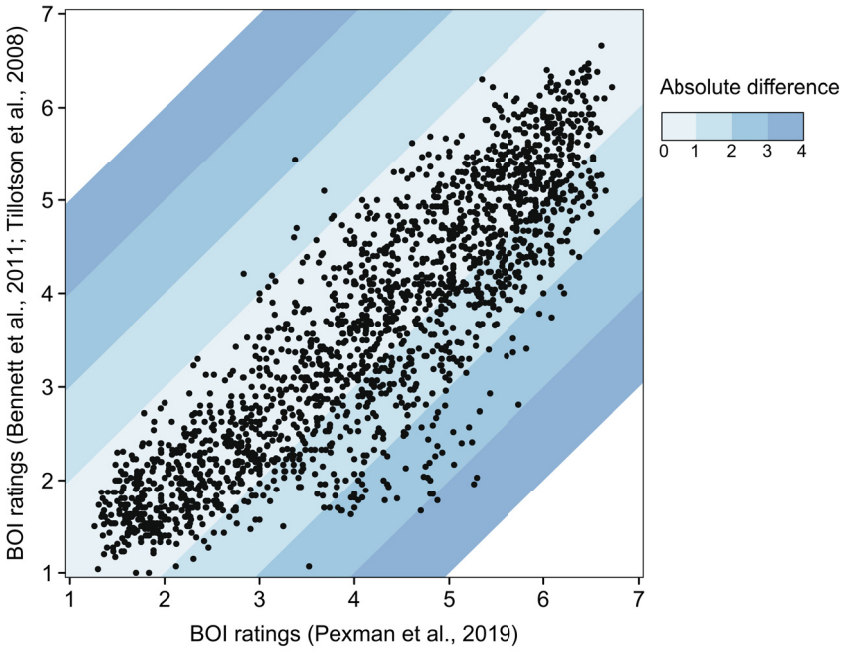
An additional and final issue is the measurement error on the ratings. Consider a word which elicits high disagreement in the population as to its BOI rating. Through random sampling, one norming study might obtain a relatively homogeneous set of responses with an average rating on one end of the scale, while another detects the disagreement and finds a midscale average. In an extreme scenario, the ratings might even end up on opposite ends of the scale. This variability certainly depends on the number of observations used to compute the average. As the question of sampling precision (see Trafimow, 2018; Trafimow & Myüz, 2019) has never been addressed in the norming literature, the extent to which this affects psycholinguistic ratings is difficult to fully evaluate. Comparing the ratings provided by different studies can nevertheless give a first glimpse at the problem. Figure 5 plots the combined ratings from Bennett et al. (2011) and Tillotson et al. (2008[25] against those from Pexman et al. (2019). A large number of items have reassuringly close ratings in the two datasets (below one unit of difference). However, several words also display the variability expected from the hypothetical cases presented above (e.g. *leopard* has a BOI rating of 1.96 in one dataset and 5.26 in the other). The midscale disagreement problem's implications

[25]The two datasets were combined because they normed different sets of words and in order to increase the overlap with Pexman et al.'s (2019) ratings.

thus extend beyond midscale items. If norming studies do not have an appropriate sample size to detect a disagreement in the ratings, a given word might be falsely detected as low or high on the scale. This raises important questions about the overall reliability of psycholinguistic norms and suggests that the typical 20 to 30 observations by word might be insufficient to obtain accurate ratings distributions.

Figure 5

Combined BOI ratings from Bennett et al. (2011) and Tillotson et al. (2008) against those provided by Pexman et al. (2019) for 1897 words in common

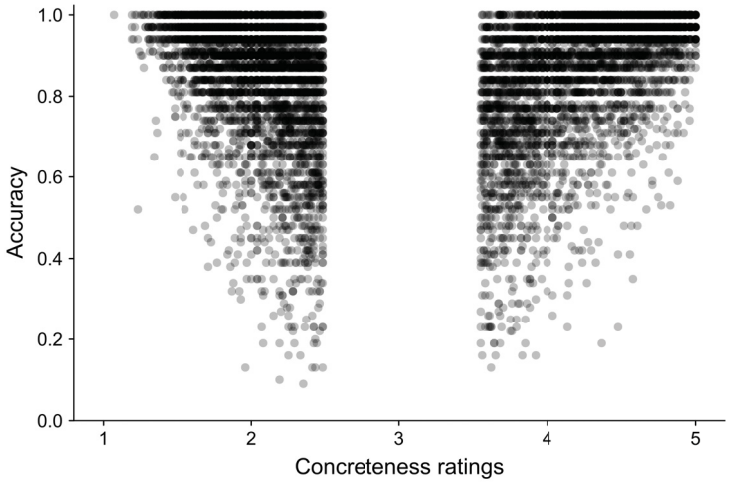


Note. The ribbons correspond to ranges of one unit of absolute difference between the ratings.

A second example in which the measurement error can be observed comes from comparing binary classifications with Likert-type ratings for the same dimension. Unfortunately, such data is not available for BOI. However, Pexman et al. (2017) conducted a semantic megastudy with over 9000 words in which participants had to classify items as concrete or abstract. Each word's category was predetermined based on Brysbaert et al.'s (2014b) 5-point concreteness ratings: items with a concreteness score in [1, 2.5] were considered abstract, and those in [3.5, 5] as concrete. Figure 6 plots each word's classification accuracy in the semantic task as a function of its concreteness rating. It is important to note that Pexman et al.'s (2017) study required speeded responses which can lead to errors. Keeping this in mind, the observation that some words had accuracies as low as 20% or less is surprising. This means that a word was considered as, e.g., 'abstract' based on the

ratings, but that more than 24 of the approximately 30 participants categorised it as 'concrete' in the semantic task. In light of the midscale disagreement problem, it appears highly unlikely that these accuracy rates truly reflect classification errors. Rather, they suggest that the concreteness ratings for some of the words were unreliable. This issue in turn affects the computation and analyses of reaction times and accuracies, with no available method to distinguish items truly associated with low accuracy rates from those that are ambiguous as to their concreteness.

Figure 6
Classification accuracies in Pexman et al.'s (2017) concrete/abstract megastudy against the concreteness ratings provided by Brysbaert et al. (2014b)



The following sections explore how the issues outlined above affect the experiments on BOI's effect in adult lexical and semantic processing. For clarity purposes, factorial design studies are reviewed separately from those using regression analyses.

3.2 Low vs. High BOI

Stimulus lists in factorial experiments are obtained through a high-low split (or dichotomisation) of the variable of interest and are matched on a number of other variables (e.g. frequency, imageability, concreteness, number of features – although often not AoA in the case of BOI). An inspection of the summary statistics for each word list in the BOI literature (Table 1) reveals that most studies must have included a large number of words towards the middle of the scale in at least one of their lists. Two exceptions are Al-Azary et al.'s (2022) experiment and Muraki and Pexman's (2021) lexical decision task (LDT) in which low- and high-BOI words appear relatively closer to the ends of the scale (but note the high SD for high-BOI words in the latter case).

Table 1
Characteristics of the low- and high-BOI lists reported by the factorial design experiments

Study	Task	*N*	Low-BOI *M*(*SD*)	High-BOI *M*(*SD*)
Hansen et al. (2012)	L & S decision	24	3.3	5.3
Siakaluk et al. (2008a)	L decision			
Siakaluk et al. (2008b)	S & S-L decision			
Xue et al. (2015)	Sentence acceptability			
Duffels (2022)	S decision	36	3.30 (0.59)	5.60 (0.47)
Hargreaves et al. (2012)	L decision			
Al-Azary et al. (2022)	S decision	45	2.56 (0.56)	5.55 (0.38)
López Zunini (2016)	L & S decision	60	2.58 (0.61)	5.00 (0.49)
Muraki et al. (2023b)	Syntactic classification	50	2.00 (0.50)	5.10 (1.00)
Muraki et al. (2023c)	S decision	50	3.00 (0.77)	5.59 (0.44)
Muraki & Pexman (2021)	L decision	50	2.39 (0.64)	5.58 (1.01)
Muraki & Pexman (2021)	Syntactic classification	50	2.04 (0.47)	5.14 (1.04)
Phillips et al. (2012)	Sentence reading	40	3.33 (0.59)	5.63 (0.44)
Tousignant & Pexman (2012)	S decision	35	3.39 (0.55)	5.67 (0.46)
Wellsby et al. (2011)	S decision	16	3.2	5.0

Note. S: Semantic; L: Lexical. *N* denotes the number of words in each list. The Low-BOI and High-BOI columns present each list's average BOI rating and its SD, when available.

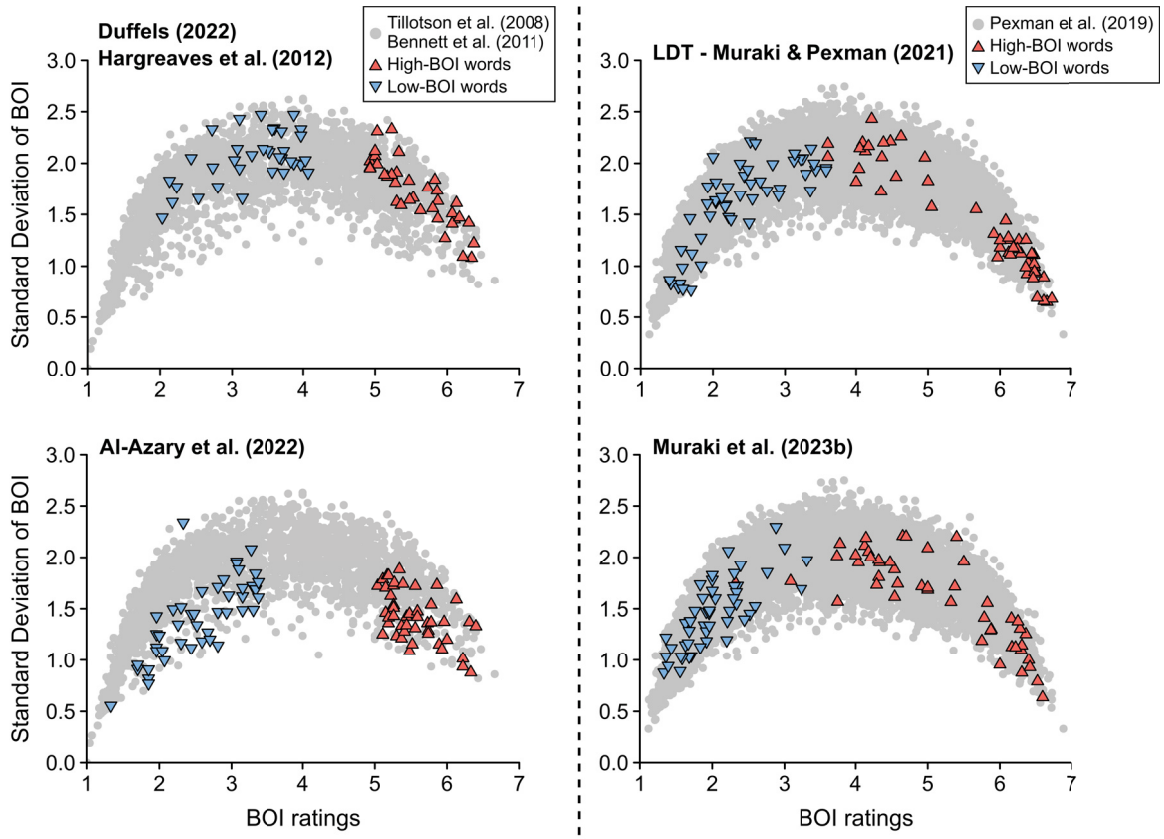
Thanks to most of these studies sharing their stimuli, it is possible to follow a similar approach to Pollock's (2018) by mapping them on the SD against means plot presented earlier in order to take a closer look at their distributions. Figure 7 depicts the stimuli used by Al-Azary et al. (2022), Hargreaves et al. (2012; Duffels, 2022), Muraki et al. (2023b), and in Muraki and Pexman's (2021) LDT, plotted on top of all the words in their reference datasets. For Hargreaves et al. (2012; Duffels, 2022), most words in their low-BOI list are found towards the middle of scale and have high SDs. Very similar distribution patterns can be observed with Phillips et al.'s (2012) and Tousignant and Pexman's (2012) stimuli, and likely with the lists of all other studies with similar summary statistics[26]. These experiments thus essentially report differences in processing between words rated high on the scale by most participants and those for which they disagreed about how to rate them – not an effect of BOI *per se*. The words used by Muraki et al. (2023b) and in Muraki and

[26]The studies by Hansen et al. (2012), Siakaluk et al. (2008a, 2008b), Xue et al. (2015) and Wellsby et al. (2011) used the ratings collected by Siakaluk et al. (2008a) which are not publicly available. Muraki et al.'s (2023c) stimuli are not provided either, but their low BOI list displays similar summary statistics (note the higher SD) to those of Hargreaves et al. (2012), Phillips et al. (2012) and Tousignant and Pexman (2012).

Pexman's (2021) LDT are relatively more representative of the entirety of the scale. The two categories are nevertheless not clearly split and several items display high disagreement – especially in the high-BOI list, which likely introduces considerable noise to the results. Finally, Al-Azary et al.'s (2022) stimuli display a much healthier distribution. Despite some low-BOI items that are close to the middle of the scale, most words have low SDs and ratings drawn from the ends of the scale.

Figure 7

BOI ratings of the stimuli used by a representative set of factorial studies against all the items in their reference datasets (left column: combined ratings from Bennett et al., 2011, and Tillotson et al., 2008; right column: Pexman et al., 2019)

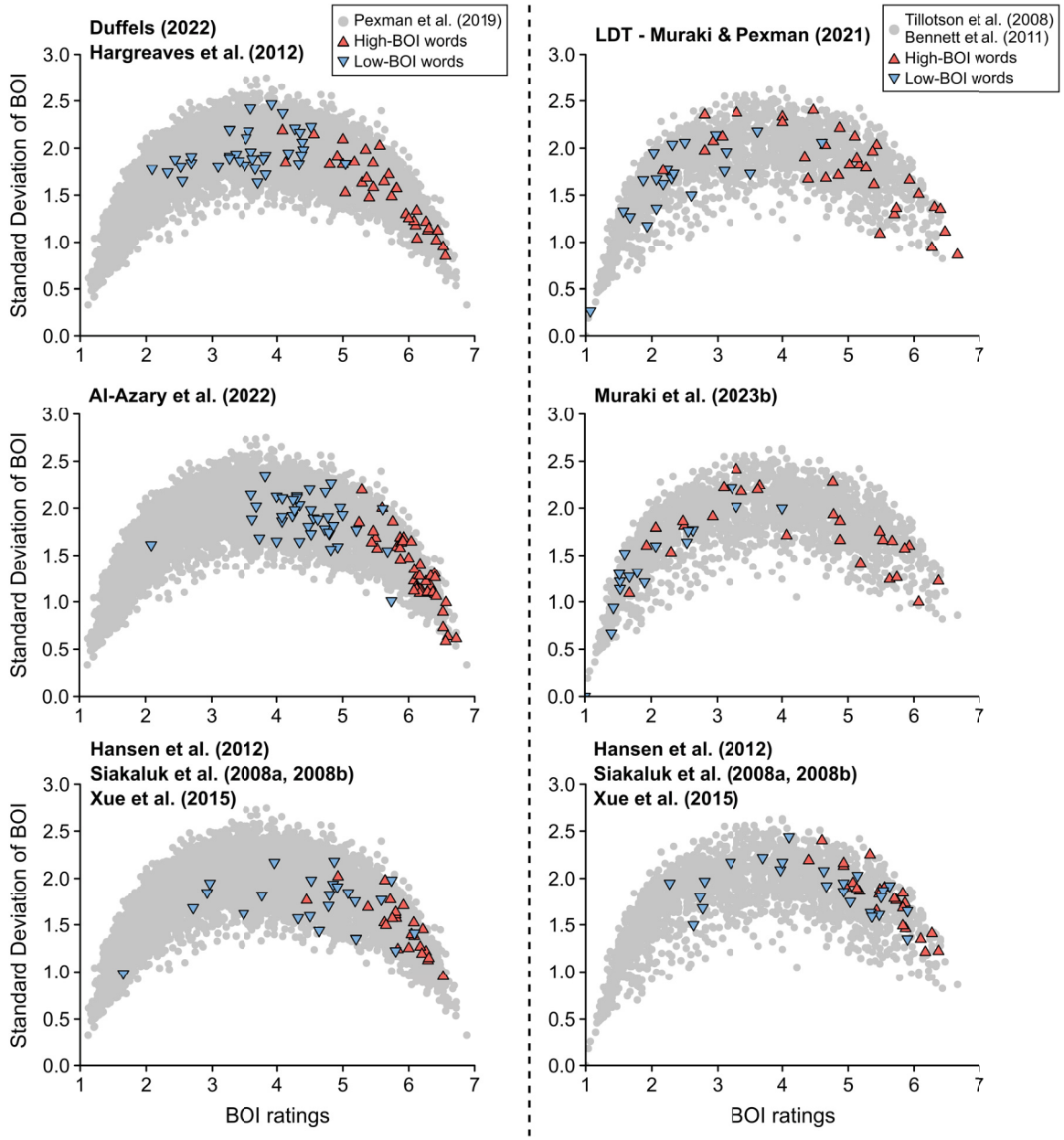


Examining the between-study rating variability for these stimulus sets raises further concerns about the validity of factorial experiments. As can be seen in Figure 8, using different BOI datasets than those from which the stimuli were originally drawn sometimes leads to alarming overlap between the low- and high-BOI lists. For most of these studies, part of the words in one of their lists could have just as well been classified as belonging to the other if the ratings had been drawn from a different dataset. The items in Al-Azary et al.'s (2022) study that were mostly found towards the ends of the scale are also heavily affected, with almost the entirety of the low-BOI words found

in the middle of the scale or even on its high end in Pexman et al.'s (2019) ratings. Overall, it is thus difficult to determine what these experiments truly compare and their results cannot be reliably attributed to an effect of BOI.

Figure 8

BOI ratings of the stimuli used by a representative set of factorial studies as found in a different dataset than that which was originally used (left column: Pexman et al., 2019; right column: combined ratings from Bennett et al., 2011, and Tillotson et al., 2008)



3.3 Regression studies

Regression analyses have been described as a statistically much more robust method to assess a variable's influence in word processing (Balota et al., 2012; Brysbaert et al., 2014a; Brysbaert et al. 2016; Keuleers & Balota, 2015). In the case of subjective variables such as BOI, however, their results remain highly sensitive to the choice of items over which the analysis is performed. Similar to most factorial design experiments reviewed above, several studies using regression models have assessed BOI's effect on words which are not representative of the entirety of the variable's scale (Bennet et al., 2011; Hargreaves & Pexman, 2014; Newcombe et al., 2012; Taikh et al., 2015; Yap et al., 2012). Critically, their samples mostly span from one end of the scale to approximately the middle (Figure 9[27]), indicating that their results essentially capture processing differences between words for which raters disagreed about their judgements and those for which they agreed on only one extreme. In the absence of words drawn from the opposite end of the scale, no reliable inferences can thus be made about BOI's role in word processing based on these analyses.

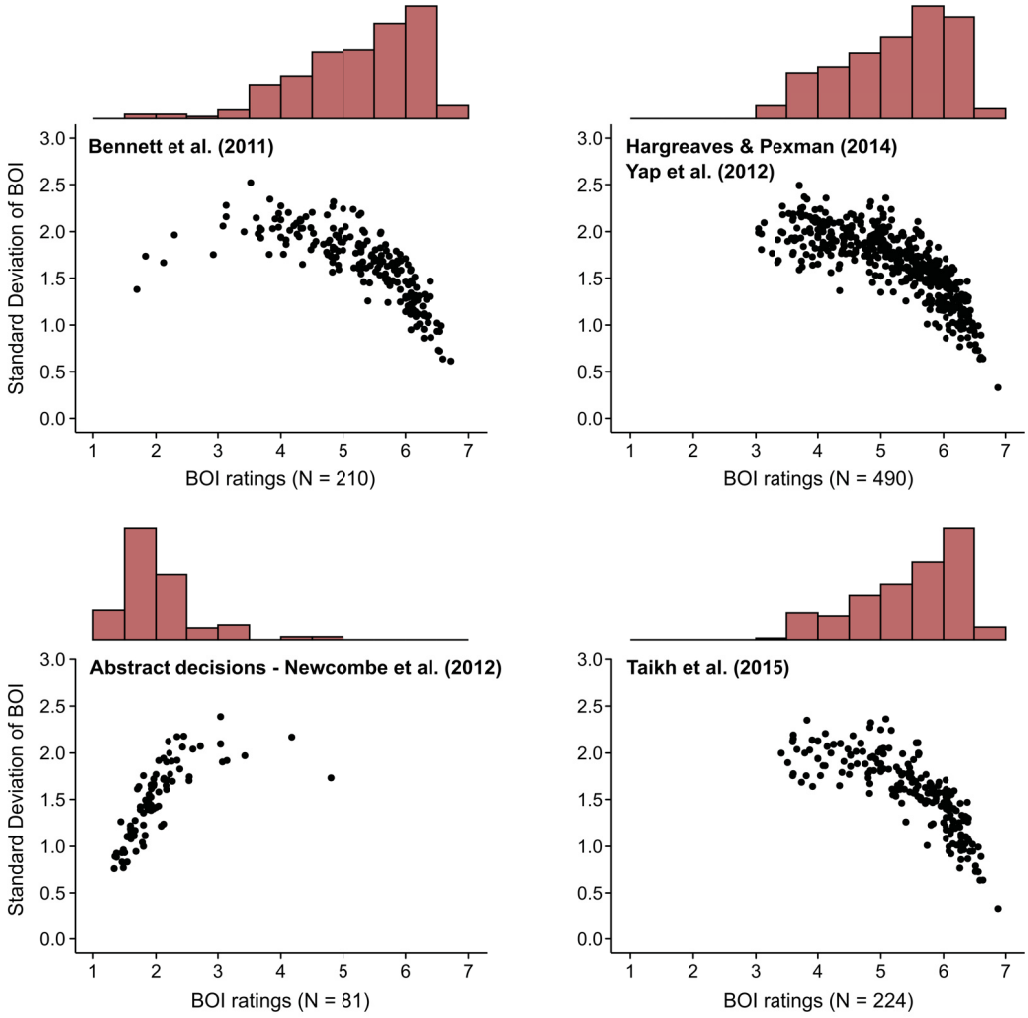
To our knowledge, eight regression studies assessing BOI's effect have included words whose ratings are distributed across the entire scale (Alonso et al., 2018; Heard et al., 2019; Juhasz et al., 2011; Lalancette et al., 2024; Newcombe et al., 2012; Pexman et al., 2019; Tillotson et al., 2008; de Zubicaray et al., 2023[28]). With the exception of Pexman et al. (2019) and de Zubicaray et al. (2023), these studies have assumed a linear effect of BOI in their models – as it is often the case with studies on subjective variables – and whether nonlinearities were considered is unclear. As was argued earlier and as Pexman et al.'s (2019) findings of a quadratic effect suggest (see Section 2.4), midscale words are susceptible to exhibit differences in processing relative to those at the ends of the scale due to confounded variables. Additionally, the midscale effect could vary from one study to another depending on the variables added to the model and on the characteristics of the sampled words. Assuming linearity in such cases makes the models highly prone to misspecification errors and can lead to unreliable estimates (Buja et al., 2019), especially when a relatively large number of midscale words are included in the analysis. It is thus possible that these studies suffer from statistical biases driven by a midscale effect.

[27] It should be noted that the ratings presented in Figure 8 are taken from Pexman et al. (2019), which were not available at the time of these experiments. The norms provided by Tillotson et al. (2008) and Bennett et al. (2011) often reveal relatively more spread-out distributions for the same items. As discussed earlier, however, this variability in ratings between norming studies is also indicative of measurement error and rater disagreement.

[28] de Zubicaray et al. (2023) present six analyses on megastudy data for different types of tasks. Five of them included BOI ratings from the entire range of the variable. However, the items used for their analysis of picture naming latencies (study 6) were predominantly drawn from the middle and high end of the BOI scale as in the data presented in Figure 9.

Figure 9

BOI ratings (Pexman et al., 2019) of the words included in the analyses of Bennett et al., (2011), Hargreaves & Pexman (2014; Yap et al., 2012), Newcombe et al. (2012) and Taikh et al. (2015).



Note. The histograms in the top margin of each plot represent each sample's frequency distribution of BOI ratings.

In conclusion, the midscale disagreement problem has important conceptual and statistical implications which affect a large number of experiments on the BOI effect – and likely on other subjective variables. In light of the concerns raised here, the most reliable results are provided by Pexman et al. (2019) and de Zubicaray et al. (2023) as their analyses were performed on a large number of words, drawn from the entire BOI scale, and because they account for potential nonlinearities resulting from a midscale effect. We would like to stress that our aim is by no means to criticise the reviewed studies. Such an analysis could not have been carried out at the time of the

experiments as large-scale and overlapping datasets have only recently been available. On the contrary, the BOI literature offers a particularly rich case study for exploring the problems inherent to Likert-type ratings thanks to the availability of multiple datasets and to the transparency of experiments regarding their stimuli. The fact that a large portion of the literature on this variable suffers from the midscale disagreement problem shows that it should be taken seriously and investigated throughout the field if any reliable conclusions are to be drawn about the mechanisms underlying word processing.

3.4 The present study

The issues raised here and by Pollock (2018) are fundamentally based on the observation that words with midscale average ratings tend to have high SDs. Although researchers should undoubtedly avoid using these items (or control for their effects through adequate statistical modelling), their bounds remain rather vague and difficult to pin down. Which ranges of means and SDs are likely indicative of high disagreement? Should all words with high SDs be avoided, or only those towards the middle of the scale? These questions are difficult to tackle based only on the summary statistics provided by datasets and require a more detailed analysis of the responses underlying them. We here present new BOI ratings for a set of 1019 French words and exploratory analyses based on their raw rating distributions. In order to have a representative sample of observations, each word was rated by a large number of participants, thus also minimising the variability problem discussed in the introduction. It should be noted that data collection started using a pen and paper (*print* hereafter) format. Due to the COVID-19 pandemic and the ensuing lockdown, the questionnaire was later transcribed to online form. This nevertheless enabled us to obtain ratings from a much more diverse sample of participants.

3.5 Methods

3.5.1 Participants

A first group of 442 participants (266 female, 170 male, 5 'other'; Age: $M = 22.14$, $SD = 4.46$) were recruited on the University of Toulouse Jean Jaurès campus and were given the print version of the questionnaire. An additional 648 participants (432 female, 206 male, 5 'other', 5 N/A; Age: $M = 25.99$, $SD = 9.34$) completed the online version. These were mainly university students across France recruited through social media platforms. All participants were over 18

years old and gave their consent at the beginning of the experiment. The experimental protocol was approved by the ethics committee of the University of Toulouse

3.5.2 Materials

The stimuli consisted of 1019 French mostly concrete nouns. An initial list of 720 words were chosen so as to maximise the overlap with other lexical (Gimenes & New, 2016; New et al., 2004, 2007), semantic (Bonin et al., 2011, 2018; Chalard et al., 2003; Desrochers & Thompson, 2009; Ferrand et al., 2008; Lachaud, 2007) and megastudy (Ferrand et al., 2018) datasets in French. An additional 299 nouns mainly referring to manipulable objects were finally added to obtain a larger dataset for use in future studies.

The instructions for the BOI ratings (provided in Appendix A) were a modified version of those used by Tillotson et al. (2008). Participants were asked to rate the ease with which the human body can physically and directly interact with what each word refers to. The ratings were given on a 0-6 scale. A rating of 0 represented an impossibility of interaction. A value of 1 meant that it is very difficult for the human body to interact with the object and a value of 6 meant that it can very easily do so. Individual labels were used for each choice (0 - impossible, 1 - very difficult, 2 - difficult, 3 - somewhat difficult, 4 - somewhat easy, 5 - easy, 6 - very easy). Participants were further asked to interpret ambiguous words, when possible, as physical objects.

In both versions of the questionnaire and in order to limit the study length to approximately 10 minutes, each participant was presented 90 words randomly sampled from the list of 1019 words[29]. The print questionnaire was a ten-page A5 booklet. Its cover page was dedicated to demographic information (age, gender, education level and domain, handedness, whether French is their native language and the age at which they learned it otherwise, if they have a known language disorder). The full instructions were presented on the third page, while the stimuli were shown from page 5 onwards with a corresponding horizontal rating scale next to each one. The scale labels were only included on the first scale of each page. Each page consisted of the main sentence of the instructions presented at the top, followed by 16 words to rate (the last page contained 10 words). The online version was designed on Qualtrics. The same full instructions were first shown on a separate page before the rating task, and a second time at the top of the page which also included the list of 90 words along with their corresponding rating scales. The words were divided into 3 consecutive blocks of 30 words. A text input box was presented after every block for participants

[29]One of the words in the lists of four participants was duplicated in the print version. Only the first rating was kept. In the online version, 14 participants were presented with 80 words and 2 were presented with 84 words due to a technical error.

to report any unknown words that they might have encountered and the scale labels were repeated for the next block.

The target of the experiment was to collect approximately 35 ratings by word for the print version and 45 ratings by word for the online version of the questionnaire. In order to achieve this, words for which the target was reached were incrementally removed from the sampling pool.

3.5.3 Procedure

For the print version, participants who accepted to take part in the study were given a consent form. Upon completion, the experimenter orally presented the booklets and rating instructions before handing the questionnaires to the participants. When multiple participants were present, the experimenter further added that they should rate the words individually, without influencing each other's ratings. In the online version, participants were similarly first asked to give their consent, which was followed by the demographic questionnaire. They were then presented with the full instructions and had to confirm that they carefully read them in order to start the rating task.

3.5.4 Data analysis and availability

All data wrangling and analyses were performed with R (v4.1.2, R Core Team, 2021) and through the RStudio interface (v2022.7.0.548, RStudio Team, 2022). All trial-level data (raw and pre-processed), summary statistics and the script used for the analyses are available in open access at this paper's Open Science Framework repository (https://osf.io/9murh).

3.6 Results

3.6.1 Data cleaning

Several criteria were used to detect invalid responses and to clean the data. These steps were applied to the combined ratings from the online and print questionnaires. First, participants who rated less than half of the words ($N = 11$) and who reported having learnt French after the age of two ($N = 40$) were removed from the analysis. To detect inattentive/careless participants, an analysis based on the inter-item standard deviation (ISD; Marjanovic et al., 2015; see also Curran, 2016) was preferred over long-string analysis as multiple words in an experimental list could refer to high BOI objects due to random sampling. ISD is simply the standard deviation of each participant's responses, with small values indicating low variation in their ratings (0 for fully uniform

responses). We removed the data of 12 participants who had an ISD lower than 3 standard deviations from the average ISD of the group ($M = 1.86$, $SD = .39$). Participants with a person-total correlation (Curran, 2016) below .20 were further dropped from the analysis ($N = 28$). Finally, we performed an item-level outlier screening and removed a total of 820 observations falling $\pm$ 3 standard deviations from the mean rating of their respective words. The results reported below are for the remaining 999 eligible participants who provided 87,414 valid ratings. There were 1,111 omitted responses and 148 entries for unknown words in the online questionnaire, and 281 omitted responses in the print questionnaire. Each word was rated by an average of 35.9 ($SD = 1.2$, $Min = 30$, $Max = 39$) participants in the print version and 49.9 ($SD = 2.65$, $Min = 41$, $Max = 60$) participants in the online version. Overall, the average number of observations for each word was 85.8 ($SD = 2.84$, $Min = 75$, $Max = 96$).

3.6.2 Reliability

The ratings from the print and online versions of the questionnaire were highly correlated ($r = .96$, $p < .001$), thus pointing to an equivalence between the two formats. Internal reliability was assessed for both versions separately and for the combined dataset through three methods which overall point to a good reliability for the present norms (Table 2). First, the split-half reliabilities were computed by averaging the Spearman-Brown corrected correlations between the mean BOI ratings over two randomly split halves (1000 iterations). Following Desrochers and Thompson (2009), we also computed the mean average absolute z scores for the three versions. These scores are computed by first standardising the ratings for each word separately and then averaging each participant's obtained values. It represents, for each participant and in standard units, the average difference between their ratings and that of the group. The mean of all participants' averages is thus an indicator of how much, on average, their ratings differed from the group. As can be seen in Table 3, the ratings provided by the participants were on average below 1 standard deviation away from the group mean ratings. Finally, person-total correlations were computed over all participants and averaged, thus indicating how much, on average, the ratings of participants correlated with the corrected group ratings. Again, the results were consistent across the datasets and similar to those obtained by Desrochers and Thompson (2009) for their ratings.

The validity of the combined ratings was further assessed by computing their correlations with the available datasets in other languages. The words in the present study's list were translated into English and matched against the items in English norms (Bennett et al., 2011; Pexman et al., 2019; Tillotson et al., 2008) and the English translations of those in one Spanish (Alonso et al., 2018) and one Russian dataset (Bonin et al., 2013). When duplicate items were present in the target set of common words, their mean BOI rating was computed before performing the

Table 2
Results of the internal reliability analyses

Reliability metric	Print	Online	Combined dataset
Split-half reliability	.96 (.00)	.97 (.00)	.98 (.00)
Mean average absolute Z score	.82 (.15)	.81 (.19)	.82 (.17)
Person-total correlation	.67 (.12)	.71 (.12)	.69 (.12)

Note. Standard deviations are reported in parentheses.

correlations ($N = 28$, Alonso et al., 2018; $N = 7$, Bennett et al., 2011). The results presented in Table 4 overall indicate fairly large correlations between the present ratings and the target norms, which are consistent with previous findings in the literature (e.g. Alonso et al., 2018; Pexman et al., 2019).

Table 3
Spearman correlation coefficients between the current ratings and those from other available datasets

Ratings	Language	N	r
Alonso et al. (2018)	Spanish	256	.85
Bennett et al. (2011)	English	200	.80
Bonin et al. (2013)	Russian	210	.84
Pexman et al. (2019)	English	846	.80
Tillotson et al. (2008)	English	408	.75

Note. All $ps < .001$

3.6.3 Descriptive analysis

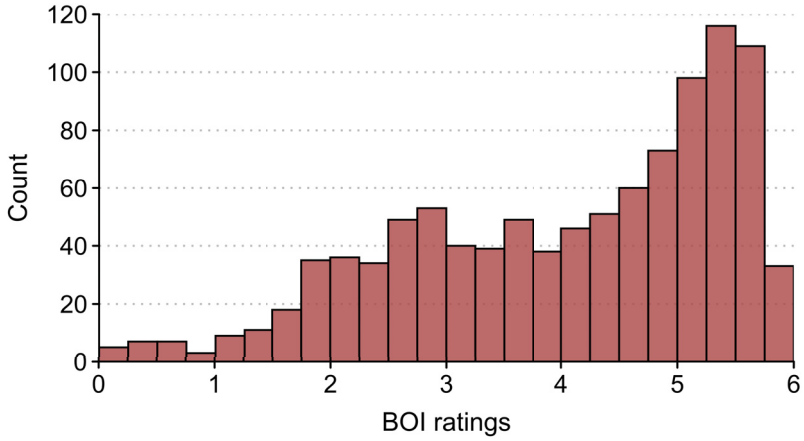
Descriptive statistics for the combined ratings are summarised in Table 4 and their frequency distribution is presented in Figure 10.

The inverted-U relationship between the average ratings and their standard deviations (SD) discussed in the introduction was also found for the present ratings and is shown in the centre plot of Figure 11. As can be seen in the examples of item-level response distributions provided in the plot's margins, words with an average rating close to 3 have highly varying response patterns. Except for those with an SD of approximately 1.5 or below (e.g. *otter*, $M = 2.77$, $SD = 1.48$), they typically display little to no interrater agreement and have multimodal or near-uniform rating distributions (e.g. *drink*, $M = 2.82$, $SD = 2.74$; *alarm*, $M = 2.92$, $SD = 2.10$; *monument*, $M = 2.89$,

Table 4
Descriptive statistics for the present BOI ratings (N = 1019) and for the metrics used to assess their reliability. The absolute difference refers to the absolute value of the difference between each word's BOI rating and its trimmed mean

	M	*SD*	*Min*	1st *Q*	*Mdn*	3rd *Q*	*Max*	Skewness	Kurtosis
BOI ratings	4.06	1.38	0.15	2.91	4.41	5.26	5.92	– 0.64	– 0.59
Trimmed mean	4.14	1.93	0.11	2.07	5.29	5.56	5.92	– 0.91	– 0.88
Absolute difference	0.66	0.55	0	0.20	0.51	0.98	2.58	0.83	– 0.10
Interrater agreement	0.76	0.16	0.43	0.62	0.77	0.92	1	– 0.18	– 1.25

Figure 10
Histogram of the present BOI ratings (N = 1019)

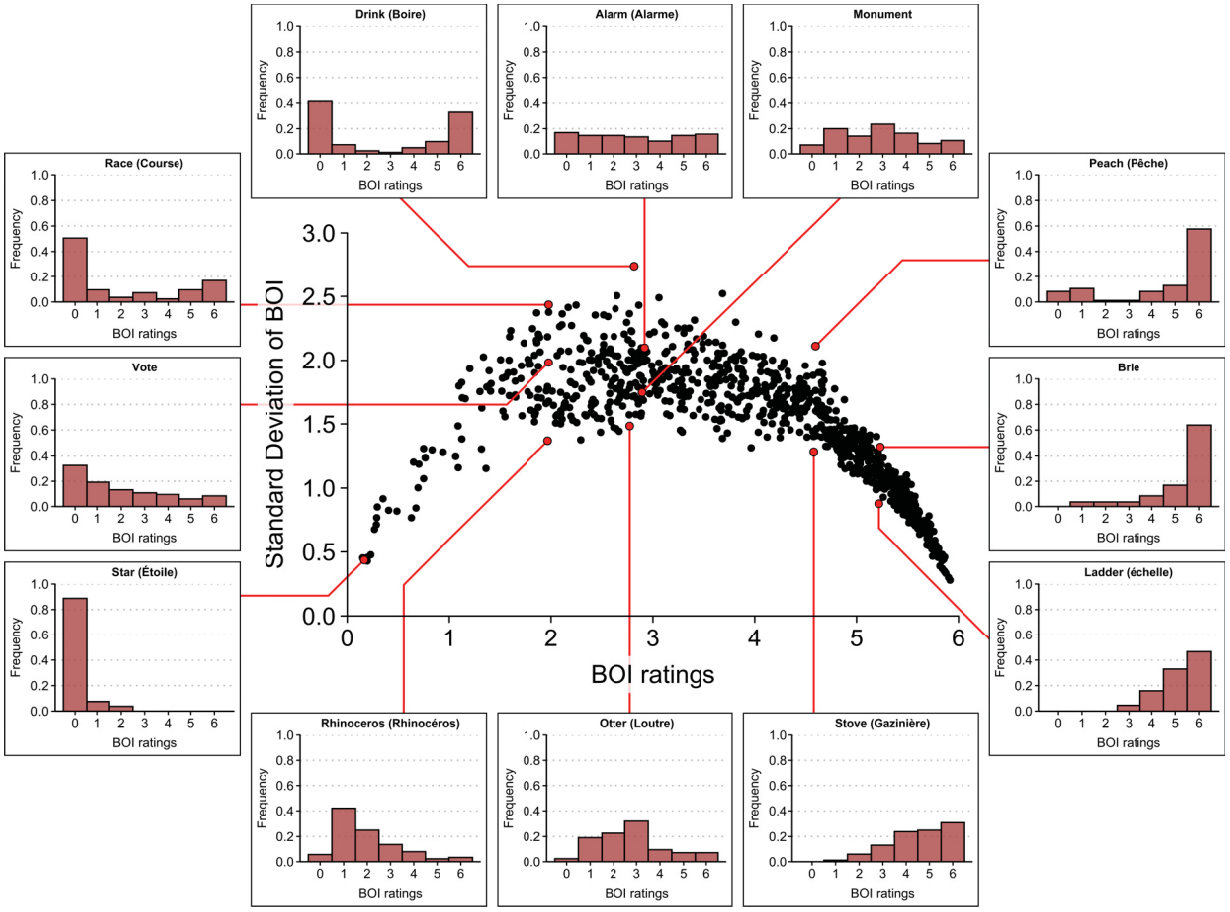


SD = 1.75). As the average moves away from 3, the judgments start to cluster at one end of the scale. For an average rating between approximately 1 and 5, higher SDs (on the "upper arc" of the centre plot) typically correspond to J-shaped distributions (e.g. *race*, *M* = 1.98, *SD* = 2.43; *peach*, *M* = 4.60, *SD* = 2.11) while lower ones to heavy-tailed distributions (e.g. *vote*, *M* = 1.96, *SD* = 1.98). Words with the lowest SDs ("lower arc" of the centre plot) generally have more consistent underlying ratings (e.g. *rhinoceros*, *M* = 1.97, *SD* = 1.37; *stove*, *M* = 4.58, *SD* = 1.28). The strongest agreement in judgements is found for words with an average rating of approximately below 1 or above 5 (e.g. *ladder*, *M* = 5.22, *SD* = 0.88; *brie*, *M* = 5.23, *SD* = 1.32; *star*, *M* = 0.15, *SD* = 0.45).

These observations clearly show that the SD is not a robust indicator of interrater agreement – rather disagreement – as its interpretation dependents on the average rating. Indeed, words with similar SDs but different means can display drastically different distributions (e.g. *alarm*, *M* = 2.92, *SD* = 2.10; *peach*, *M* = 4.60, *SD* = 2.11). The descriptive analysis also concurs with the issues

Figure 11

Standard deviation as a function of the average BOI ratings in the present study (centre) and examples of item-level rating distributions



raised in the introduction. To a large extent, agreement among raters seems to gradually decrease as the average approaches the midpoint of the scale, except for words with relatively low SDs (bottom panel of Figure 11). Similarly, the average rating becomes decreasingly representative of how participants responded the closer it gets to the middle for most items. In the following sections, we assess the item-level interrater agreement and the distance between the majority of responses to the average rating to get a more comprehensive look at how they relate to the traditional summary statistics.

3.6.4 Interrater agreement

Several measures of interrater agreement – or consensus metrics – for Likert-type scales exist in the literature (for a review, see O'Neill, 2017. See also Claveria, 2021; Abdal Rahem & Darrah,

2018; Tastle & Wierman, 2007). However, these present a number of important disadvantages. Chief among them is that they often fail to give a satisfactory value across all of the response profiles described above and that they are difficult to interpret. For the current descriptive purposes, we were interested in a straightforward measure of the extent to which participants' responses are aggregated.

Agreement was defined as the highest proportion of responses in any range of 3 consecutive BOI units (i.e. among the proportions of ratings falling in [0, 2], [1, 3], [2, 4], [3, 5] and [4, 6]). We chose 3 scale units because they cover conceptually consistent response options, as well as to capture the naturally higher dispersion in midscale words' rating distributions. Descriptive statistics for the agreement scores are presented in Table 4. We also identified words with multimodal response patterns based on the kernel density distribution of their ratings (bandwidth = .5) using the *peaks* function from the IDPmisc R package (version 1.1.20; Locher, 2020) and some additional constraints. A word's distribution was labelled as multimodal if its density distribution had at least two peaks, separated by 3 BOI units or more, and if the height of the peaks was superior to half of the highest one's. Among the 1019 rated words, 89 were detected as having a multimodal distribution. Note that this method led to items with approximately uniform distributions to also be detected as having a multimodal distribution.

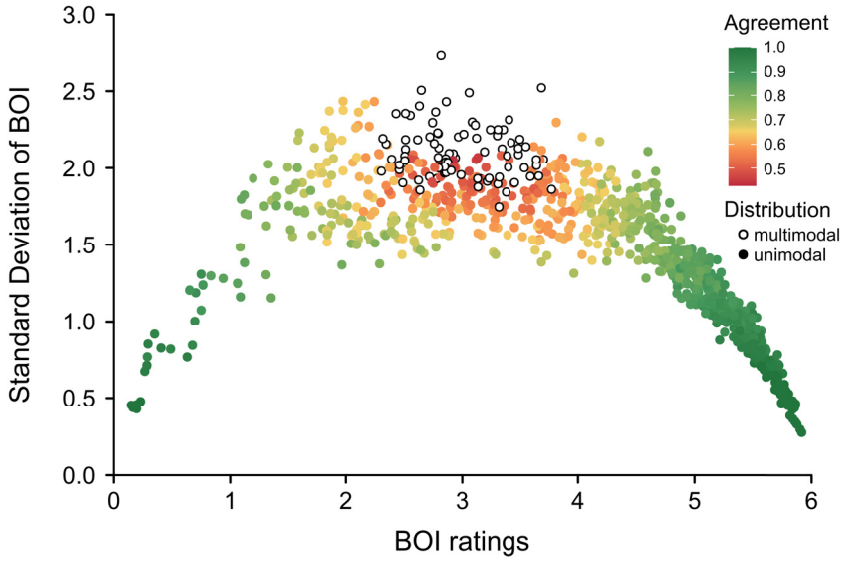
The results presented in Figure 12 confirm and help to generalise the previous observations. Words with an average rating between approximately 2 and 4 and with an SD above 1.5 generally have an agreement score below .65[30] or display a multimodal distribution. In the former case, there were thus less than 65% of participants who responded within 3 BOI units. Some exceptions can be found on the left half of the plot which have agreement scores above .65 and an SD slightly above 1.5. These emerge as a result of how the scale was labelled (i.e. 0 – 'impossible', followed by 1 – 'very difficult' to 6 – 'very easy'). They mostly refer to animals which were rated as difficult – but not impossible – to interact with by most participants (e.g. *owl*, $M = 2.57$, $SD = 1.67$, $Agreement = .70$; *penguin*, $M = 2.40$, $SD = 1.67$, $Agreement = .74$).

3.6.5 Trimmed means

The agreement measure used above is not informative as to the average rating's reliability; it only captures the aggregation of judgements. As was previously shown, many words have heavy-tailed or J-shaped distributions with most of their data clustered at one end of the scale. These can have relatively high agreement scores, but their average rating is drawn towards the middle of the scale by extreme values and is thus not representative of the underlying distribution.

[30]This value is only chosen as a reference point to facilitate the interpretation with respect to the average rating. It is not intended as a threshold for an acceptable agreement rate.

Figure 12
Standard deviation as a function of the average BOI ratings, along with item-level interrater agreement scores and type of rating distribution.

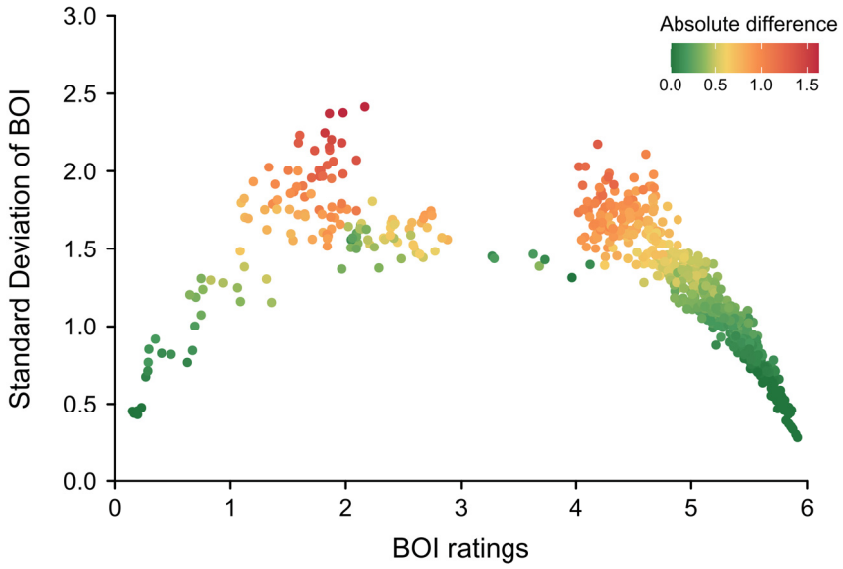


In order to better assess the distance of the average rating to the bulk of the data, we adopted a similar approach to computing a trimmed mean. For each word, we averaged the ratings falling in the same interval as the one which was used for the agreement score (i.e. within the 3 consecutive BOI units with the highest proportion of responses). We then computed the absolute difference between this value (*trimmed mean* hereafter) and the overall average rating. Descriptive statistics for both the trimmed means and the absolute differences can be found in Table 4. Figure 13 maps the differences on the SDs against means plot. For better readability, only the words with an agreement score above .65 are presented. In line with the previous descriptive analysis, the plot reveals that the average ratings are most representative of the underlying data at the two ends of the scale and that they are increasingly skewed as they approach its middle. For most words with an SD above approximately 1.5, the trimmed mean is found either within [0, 1] or [5, 6].

3.7 Discussion

The present chapter's goal was to explore the implications of the midscale disagreement problem for subjective norms and for the studies using them. We used the literature on Body-Object Interaction (BOI) ratings as a case study for our analyses as a large number of both factorial and regression studies have been conducted on the variable's effect and because overlapping datasets are available. Following Pollock (2018), we showed that the standard deviation (SD) of BOI ratings

Figure 13
Standard deviation as a function of the average BOI ratings for words with an agreement score above .65 and absolute differences between the trimmed means and the average ratings



(Pexman et al., 2019) display a concave relationship with the average ratings. Arguing that the amount of observed deviation for midscale items can only be explained by significant disagreement among raters, we derived three important implications for the ratings. First, the average rating of a large number of midscale words is not representative of their true position on the scale. These words fall outside of the variable's continuum and their ratings can thus not be compared to other words'. Second, several factors which can affect word processing contribute to the disagreement in BOI ratings (e.g. word ambiguity, animacy). If these variables are not controlled, the use of midscale words can introduce independent effects on word processing performances which would be falsely attributed to BOI. Finally, the disagreement problem can result in significant measurement error as evidenced by the variability in ratings between norming studies. Although preliminary, our analysis suggests that the 30 observations by word typically collected by norming studies are insufficient to obtain a reliable distribution of ratings in the presence of disagreement. This observation is in stark contrast to Montamedi et al.'s (2019) recommendation of a sample size of 10 observations and calls for further investigation.

We performed a methodological review of the studies on BOI's effect to assess the extent to which they suffer from the midscale disagreement problem. We showed that factorial design studies comparing low- to high-BOI words had predominantly used midscale items in their low-BOI lists. Their results are thus not informative as to BOI's effect on word processing performances. Rather,

they capture the effect of disagreement in BOI judgements which is likely driven by confound variables. We additionally showed that some of these studies' stimuli display extensive variability in their ratings across different BOI datasets, thus further challenging their validity. Our review reveals that these limitations also apply to studies investigating the variable's effect using regression analyses. Some of the studies included words with ratings ranging from approximately the middle to only one end of the scale in their models. Similarly, the regression coefficients that they report thus reflect a relationship between processing performances and levels of disagreement in BOI judgements, not varying degrees of BOI. The remaining regression studies have used a pool of words drawn from the entirety of the BOI scale. Except for Pexman et al. (2019) and de Zubicaray et al. (2023), however, these have assumed BOI's – and other variables' – effect to be linear. In light of a potential midscale effect due to confound variables, not accounting for nonlinearities makes the models prone to misspecification errors and can produce biased estimates (Buja et al., 2019. For examples of nonlinear effects with subjective variables see Bonin et al., 2018; Kousta et al., 2011). Their results should thus be interpreted with caution. It is important to stress that these remarks are not limited to the experiments on BOI's effect and likely affect those investigating other subjective variables as well (for some examples on memory experiments, see Brainerd et al., 2021; Pollock, 2018).

Several practical conclusions can be drawn from this initial analysis for researchers using Likert-type ratings. First, midscale items with large SDs should be avoided in factorial design studies. This can prove to be difficult – even unfeasible at times – when trying to match the experimental lists across other variables (Cutler, 1981). However, including midscale items directly affects the experiment's validity. If a regression study is planned instead, particular attention should be paid to include enough items from both ends of the scale so that the variable's effect can be determined. Additionally, we strongly recommend the use of nonlinear regression methods to account for potential midscale effects which can bias the results under a linearity assumption. Finally, even though determining an adequate sample size for norming studies is beyond the scope of the current work, we strongly advise against collecting less than the typical 30 observations for each item as it highly increases the probability of missing disagreements present in the population. This can lead to falsely detect items as low or high on the scale and thus to draw false inferences.

The above recommendations can serve as general guidelines for the use of Likert-type ratings. Means and SDs nevertheless remain difficult to interpret and only provide a partial picture of which words suffer from a disagreement in judgements and to what extent. A comprehensive understanding of these issues requires a more detailed analysis of raw rating distributions and of additional metrics. As item-level responses are seldom made available by rating studies, we collected new BOI ratings for a set of 1019 French words. Data collection was initially carried out

through a pen and paper format and continued online due to the COVID-19 pandemic. The average ratings obtained through the two formats were highly correlated ($r = .96$, $p < .001$) which led us to combine the data for the final ratings. Their internal reliability was assessed through standard methods (person-total correlation, mean average z-score, split-half reliability), as well as a comparison with other BOI datasets. All analyses pointed to an overall good reliability of our ratings. Additionally, each word was rated by a large number of participants ($M = 85.78$, $SD = 2.84$, $Min = 75$, $Max = 96$) to obtain a representative distribution of responses.

As with Pexman et al.'s (2019) ratings, the words in our dataset displayed an inverted-U relationship between their SDs and their average ratings. We performed a descriptive analysis of item-level rating distributions to explore the possible response profiles and their relation to the summary statistics. Our results showed that the SD is not a robust index of agreement among raters because the information it conveys about the underlying responses changes based on the word's position on the scale. This is largely due to the scale's bounded nature which results in varying ranges of possible SDs as a function of the average (Akiyama et al., 2016). We additionally found that most midscale words exhibited either multimodal or near-uniform distributions and thus that their average ratings are uninformative about the underlying data. Only a small number of words with SDs close to 1.5 displayed relatively consistent responses. Moving away from the middle, responses were increasingly clustered towards one end of the scale. These typically displayed J-shaped or heavy-tailed distributions, except for words with the lowest SDs for which the distributions resembled skewed Gaussians. Unsurprisingly, words with the highest aggregation of judgements were found close to the ends of the scale.

These observations led us to perform further exploratory analyses to detect multimodal distributions and to assess the interrater agreement on a larger scale. We defined agreement as the highest proportion of responses obtained for any 3 consecutive BOI units. Confirming and refining the previous analyses, we found that words with an average rating within approximately 1 unit of the middle of the scale and an SD above 1.5 mostly had either multimodal response distributions or low agreement scores. Irrespective of their SD, words closer to the ends of the scale had increasing agreement scores. As the previous descriptive analysis revealed, some words display J-shaped or heavy-tailed distributions. These can have relatively high agreement scores but their average ratings are skewed towards the middle by extreme values. To assess the extent to which the average rating is a reliable reflection of the underlying responses, we computed its distance to a trimmed mean based on the same 3-unit interval used for the agreement scores. For most words with an agreement score above .65, those with an SD approximately above 1.5 had a trimmed mean falling either in the first or the last interval of the scale. The majority of responses for most words were thus usually found at the ends of the scale, making the average rating an increasingly biased estimate of central

tendency the further away it gets from them and the higher its SD.

To summarise, as the average rating moves from the ends of the scale towards its middle, what it represents for most words gradually shifts from being *low* or *high* on the dimension to being *undefined* due to increasing disagreement among raters. The SD, on the other hand, mainly captures the skewness of the average rating relative to the majority of the judgements. Its ranges nevertheless change as a function of the average and higher SDs do not necessarily indicate higher disagreement as some studies have assumed (see, e.g., Brainerd et al., 2021; Strik Lievers et al., 2021; Winter et al., 2024). Indeed, some few words towards the middle of the scale display high interrater agreement despite their relatively higher SDs compared to the ends of the scale. Their small number nevertheless likely makes them negligible for any practical purposes. These observations overall highlight that the reliability of the ratings has to be assessed as a joint function of both the average and the SD. More generally, and as Pollock (2018) also pointed out, such variables cannot be taken to represent a continuous and linear theoretical dimension and should be treated accordingly. It is important to note that these observations cannot be directly generalised to other types of scales such as bipolar (e.g. valence. Brainerd et al., 2021; Pollock, 2018) or numerical ones (e.g. age of acquisition. Xu et al., 2022). Indeed, the ratings derived from these display different relationships with their SDs and should be analysed separately.

Likert-type scales are a crucial tool for researchers tackling questions about lexical, perceptual and conceptual processing. Given their predominant use and the increasing effort and budget dedicated to their collection (Hollis & Westbury, 2018), the issues raised here and by Pollock (2018) call for more attention to their methodological and statistical underpinnings which have hitherto been largely ignored. Our analyses provide a first step in this direction by enabling a more informed reading of the standard summary statistics and a more appropriate use of the ratings. We hope that these results will prove useful as general guidelines to conduct future studies and to reassess the validity of previous findings in the literature. Several critical questions about psycholinguistic ratings remain to be addressed. One of the most important is arguably the number of observations necessary to obtain reliable ratings, which directly affects the validity of the experiments using them. Testing and establishing clear guidelines for participant screening and data cleaning would also greatly benefit the field by reducing the noise in the measurements. Finally, and in light of the limitations that average ratings present, a larger discussion about the methods used to norm stimuli appears necessary. Future research should investigate in more detail the validity and limits of alternative methods for deriving ratings (e.g. Hollis, 2018; Taylor et al., 2022). To facilitate inquiries into the subject, we urge researchers to make their trial-level data from rating studies openly accessible.

3.8 Conclusion

In the continuity of Pollock's (2018) observations, our analyses strongly support the idea that Likert-type unipolar ratings should be interpreted as binary classifications with varying degrees of agreement instead of reflecting a continuous dimension. The middle of the scale predominantly captures disagreement, which can both result in differences in word processing performances for midscale items and in high measurement error in the ratings. We have seen that these issues have serious implications for a large number of studies investigating BOI's effect in word processing – especially those with factorial designs. Based on the few experiments using an appropriate range of BOI ratings – and assuming no other major methodological issues – the only reliable conclusions about the variable's effect in adult word processing appear to be that it plays no role in lexical decisions and that it has a facilitatory effect in concrete/abstract semantic classifications.

Our analysis of the extant literature was partly based on the observation that some items display considerable variability in their ratings across normed datasets. These differences were interpreted as reflecting measurement error in the presence of disagreements and an insufficient number of participants to detect them. It should be pointed out, however, that the comparison between Pexman et al.'s (2019) ratings and the combined datasets of Bennett et al. (2011) and Tillotson et al. (2008) also suggest a more systematic relationship. Indeed, some items appear to have received relatively higher ratings in Pexman et al.'s (2019) study, while the reverse pattern is rarely observed. This result is difficult to interpret but suggests that a methodological difference might be partly responsible. Our analyses in the next chapter help to shed some light on this observation and highlight some examples of how norming procedures can affect the ratings (for a brief discussion of the variability in BOI ratings, see the conclusion of Chapter 4).

Chapter 4

Through the forest of manipulability ratings

We have seen in the introductory chapters that the role of the motor system in the processing and representation of manipulable objects – and of concepts more generally – has received a great deal of attention over the past three decades and that it continues to be a hotly debated topic. Similar to psycholinguistics research, a common methodological practice in this literature is to collect Likert-type manipulability ratings which aim to capture, in a broad sense, the manual motor content associated with objects. Such ratings serve stimulus selection and control, as well as correlational analyses on behavioural and neuroimaging data, which makes them crucial for experimental validity. However, researchers who want to manipulate this dimension are quickly confronted with a multitude of scales to choose from, and little information as to which is more appropriate (see also Section 2.5). Indeed, there has been overall little consensus or discussion over how to operationally define manipulability and, critically, "it is not clear whether manipulability relates to the same motor dimension across normative studies" (p. 1, Guérard et al., 2015. For a similar point, see also Nickels et al., 2022). This definitional heterogeneity complicates the interpretation of experimental findings, raises concerns about their comparability and more generally creates confusion in the literature.

One of the main hurdles in achieving a more standardised approach is arguably the difficulty in determining what subjective scales fundamentally capture (i.e. their construct validity. See Section 1.2.4). Several studies offer a conceptual discussion of the different rating instructions used for manipulability (e.g. Clarke & Lundington, 2018; Guérard et al., 2015; Heard et al., 2019; Salmon et al., 2010). These have been instrumental in pointing out the limitations of commonly used definitions and in aligning methodological practices with advancements in our understanding of object processing. However, they typically remain limited in scope and do not ultimately allow to discern among alternative instructions used for the same motor dimension across studies. The latter point requires a more detailed analysis and comparison of the data obtained with each instruction – a highly challenging task in the case of manipulability. Studies using different instructions either do not provide readily accessible data (mainly pilot studies for stimulus selection) or vary greatly in their sets of rated items, the nature of their stimuli (e.g. words, photographs, drawings) and their geographical origin – which introduces cultural and linguistic differences. This makes it impossible to systematically compare them and their respective effects in experimental tasks without acquiring large amounts of new data (for a similar point, see Guérard et al., 2015).

In the current chapter, we aim to use a different approach to assessing the validity of manipulability ratings. Namely, we propose to tackle the mentioned difficulties by leveraging the recently highlighted midscale disagreement problem in Likert-type ratings (Chapter 3. Paisios et al., 2023; Pollock, 2018) to our benefit. In addition to qualitative and correlational analyses, this enables us to analyse the rating instructions in light of the disagreements they generate in participants' responses,

and thus to detect potential ambiguities in their wording or in their interpretation. More generally, this work also serves as an overview of the different types of instructions used for – or closely related to – the assessment of manipulability. In what follows, we start by briefly presenting the historical evolution of manipulability ratings in the context of relevant theoretical developments (Section 4.1). We then extend the midscale disagreement problem to the manipulability ratings identified in the literature (Section 4.2). These two sections are used to outline the organisation and methodology of Section 4.3, which provides an in-depth discussion of the different operationalisations of manipulability. It should be noted that the current work does not intend to exhaustively list all experiments, rather to bring out the most commonly used types of instructions and to provide a first attempt at assessing their validity. To our knowledge, it nevertheless covers most – if not all – rating studies and a representative set of experiments with a pilot rating phase. We trust that this will serve as a valuable reference point for future work and help researchers make more informed decisions about how to operationalise manipulability.

4.1 A brief history of manipulability ratings

Interest in manipulability largely originates from neuropsychological investigations into category-specific semantic deficits (e.g., Warrington & Shallice, 1984) and the discussions regarding the role of sensory-motor information in the organisation of semantic memory (Allport, 1985; Warrington & Shallice, 1987). As we have seen in Section 1.1.2 and in a similar vein to the embodied accounts of knowledge (Section 1.1.3), the key insight behind these approaches was that our knowledge of the world and of its constituents is encoded within the very systems through which we perceive and engage them. As manipulable objects are to a large extent defined by their use, motor information should similarly play a central part in their recognition and representation. Manipulability ratings have been widely used to investigate this hypothesis and emerged in a context of increasing reliance on – and availability of – subjective ratings to control various stimulus characteristics (e.g. age of acquisition, familiarity, concreteness. Carroll & White, 1973; Paivio et al., 1968; Snodgrass & Vanderwart, 1980).

To our knowledge, the first manipulability-related Likert-type ratings were collected in a study by Feyereisen et al. (1988) who asked their participants to rate “the extent to which one can act upon or with the item” (p. 404). The authors were more specifically assessing the operativity of objects, which includes their manipulability (Gardner, 1973, 1974). Using a more direct approach, Howard et al. (1995) separated the concept of operativity into its different components and collected, among others, ratings for manipulability. Unfortunately, the specific instructions used in their study are not reported, but their article implies that participants were asked to simply rate the

extent to which objects are manipulable. An important precursor rating study was also conducted by Tranel et al. (1997) in order to investigate the factors underlying category-specific semantic deficits. Contrary to the two previous studies, the definition of manipulability used by the authors puts a particular emphasis on the functional use of objects. It was operationalised as "the extent to which use of human hands is necessary for this object to perform its function" (p. 1331), and several rating studies have since used the same instructions (Miklashevsky, 2018; Moreno-Martínez & Adrados, 2007; Moreno-Martínez & Montoro, 2012; Moreno-Martínez et al., 2011, 2014; Navarrete et al., 2019). Tranel et al. (1997) also report a highly related second scale that assesses whether there is a "characteristic or typical motion associated with the use of this object" (p. 1331). These instructions are conceptually similar to those presented next but appear relatively less intuitive for raters – which might explain the lack of their use in future studies.

The first publicly available and arguably most influential set of manipulability ratings was presented by Magnié et al. (2003). This study introduced an original operationalisation of manipulability (commonly referred to as pantomime ratings) in terms of whether participants could "easily mime the action usually associated with this object so that any person looking at [them] doing this action could decide which object goes with this action?" (p. 524). By emphasising the recognisability of the actions associated with objects, these instructions aim to capture their relative distinctiveness – an aspect absent from most other manipulability ratings. The authors do not directly discuss their choice of instructions but implicitly refer to the role of distinctiveness in facilitating object recognition. This operationalisation was also likely inspired by the wide use of pantomime tasks in clinical neuropsychology to assess tool-use disorders (Osiurak et al., 2021). Magnié et al.'s (2003) definition has been used as a reference point by a large number of subsequent studies and is one of the most common assessments of manipulability in the literature (Brodeur et al., 2010, 2012, 2014; Campanella et al., 2010; Mahon et al., 2007; Wang et al., 2023).

The time around Magnié et al.'s (2003) ratings is marked by an increasing recognition of manipulability's importance in object recognition, but also by a growing interest in embodied cognition (Section 1.1.3). The latter development in particular resulted in a new wave of interest for the role of experiential and modality-specific information in how knowledge is represented, and in a proliferation of ratings assessing them. Among these, several different scales have been proposed to capture the extent to which objects are associated with motor information such as body-object interaction (see Section 2.4 and Chapter 3. Siakaluk et al., 2008a), their association to action (Carota et al., 2012; Gainotti et al., 2009, 2013; Hoffman & Lambon Ralph, 2013; Magri et al., 2021; Proverbio et al., 2011; Rueschemeyer et al., 2010b), modality-specific action strengths (Binder et al., 2016; Carota et al., 2012; Dreyer et al., 2015; Lynott et al., 2020; Magri et al., 2021; Repetto et al., 2023), and variations on their manipulability (Desai et al., 2016; Fernandino et al., 2016;

Haddad et al., 2024; Howard et al., 1995; Kellenbach et al., 2003; Medler et al., 2005; Pecher et al., 2013). Importantly, however, most studies assessing these dimensions lack a justification or a discussion of their specific operationalisations and, with a few exceptions, their instructions are generally agnostic as to the exact nature of the motor information being assessed.

In parallel, seminal work on visual object processing (Section 1.3.1. E.g., Goodale & Milner, 1992; Ungerleider & Mishkin, 1982) and subsequent refinements (Section 1.3.2. E.g., Buxbaum, 2001; Buxbaum & Kalénine, 2010; Rizzolatti & Matelli, 2003) highlighted fundamental differences in object-directed actions. As we have seen, this line of investigation notably proposes that the prehension of objects based on their structural features and their learned, skilful use are subserved by different processing streams – only the latter being considered as constitutive of conceptual knowledge. We will refer to these respectively as structural manipulability (SM) and functional manipulability (FM) in the rest of the chapter. Following this distinction, Salmon et al. (2010) offered a brief critical discussion of how manipulability had been previously operationalised and noted that the available ratings did not capture the ability to grasp objects. They thus set out to conduct the first norming study to define manipulability along two separate dimensions, namely graspability and functional usage (note that the same distinction was made at around the same time in the pilot rating tasks of Rueschemeyer et al., 2010b, and Vingerhoets et al., 2009). Salmon et al.'s (2010) study highlighted important conceptual limitations with manipulability scales and largely set the stage for their refinement. One of their key contributions has notably been to explicitly argue for the necessity of assessing SM on a separate scale (i.e. graspability), which has been implemented by several studies since (Guérard et al., 2015; Iachini et al., 2014; Heard et al., 2019. See also Amsel et al., 2012; Clarke & Lundington, 2018; Díez-Álamo et al., 2018). It should nevertheless be pointed out that Salmon et al.'s (2010) choice of instructions for the two dimensions are rather counterintuitive in light of their discussion and present a number of validity issues themselves (see Sections 4.3.2.8 and 4.3.3).

A final and important recent contribution was made by Guérard et al. (2015). Discussing the diversity of definitions used for this dimension, the authors proposed Magnié et al.'s (2003) pantomime instructions as the best candidate to capture functional actions. They nevertheless recognised the limitation that this scale results in some functionally manipulable objects receiving low ratings because of their associated actions being shared by a large number of other objects (e.g. *apple*. See also Salmon et al., 2010). This led them to propose a variation of the original pantomime instructions by asking participants to rate the ease with which they could pantomime an object, "regardless of the capacity to recognise [it]" (p. 2). The modified version has been used by two of the three subsequent rating studies collecting pantomime ratings (Clarke & Lundington, 2018; Heard et al., 2019). More importantly, the authors introduced the idea that a single scale might not adequately

capture a given motor dimension (i.e. SM or FM) and argued for the use of a conjunction of them. Their study reports two scales assessing the structural (ease to grasp and ease to move), and two scales on the functional (ease to pantomime and number of actions associated with the object. For the latter, see also Lagacé et al., 2013) manipulability of objects (for a similar approach, see Heard et al., 2019).

A more thorough and data-driven discussion of the different types of instructions used to assess manipulability across the literature is presented in Section 4.3. In line with current approaches, the review is organised into three parts related to the different dimensions of manipulability. More specifically, we distinguish the instructions assessing SM, FM, and those which explicitly refer to both dimensions (mixed manipulability, MM). Before jumping into the review, we take a necessary methodological detour in the following section to introduce a key element in how the validity of the different instructions will be assessed.

4.2 The midscale disagreement problem in manipulability ratings

We have seen in the previous chapter that subjective ratings collected through unipolar Likert-type scales are affected by the *midscale disagreement problem*, i.e. that most items in the middle of the scale are a result of disparate judgements – not a consensus among participants. Building on Pollock's (2018) observation of this issue, we have tried to map the summary statistics commonly reported for such ratings (i.e. means and standard deviations – SD) in order to better understand which items are indicative of a disagreement. The exploratory analyses revealed that items with a SD approximately above 1.5 and falling in the middle third of a 7-point scale were rated by less than 65% of participants within a 3-unit interval. These items can thus be regarded as having too low of an agreement rate for their averages to be considered as representative of how raters have responded. Pollock (2018) similarly argued that a SD above 1 on a 5-point scale indicates high disagreement and that the vast majority of midscale items do not reflect a consensus among participants.

Although preliminary, these observations provide general guidelines for the interpretation of Likert-type unipolar ratings. They suggest, for any similar scale, that an average in the middle third of the scale (*midscale* hereafter) and a SD roughly above half of the theoretical upper bound[31] are

[31]The SD of a random variable which takes values in $[a,b]$ (e.g. ratings on a Likert-type scale that stretches from a to b) has an upper bound of $\frac{(b-a)}{2}$ (Popoviciu, 1965). For 5 and 7-point scales, the upper bound for the SD respectively corresponds to 2 and 3. Note that some studies report items with SDs above this threshold. The sample standard deviations can indeed slightly exceed this upper bound which is computed for the population standard deviation.

symptomatic of a strong disparity in the underlying ratings. It is important to note, however, that these thresholds are necessary simplifications for our current purposes. Indeed, the 65% agreement threshold is rather conservative and items close but not directly in the middle third of the scale can still display high disagreement – especially when their SD is high. Additionally, the measurement error in Likert-type ratings remains an open question in the literature (see Chapter 3) and can likely result in some disagreed-upon items not being detected in the middle of the scale. These can nevertheless be expected to have ratings that are close to the middle third of the scale.

Most manipulability scales are unipolar (except for the numerical 'number of actions' scale, Guérard et al., 2015; Lagacé et al., 2013; Heard et al., 2019) and vary between 5 and 8-points. They are thus comparable to those analysed by Pollock (2018) and in Chapter 3. This enables us to use the guidelines presented above to screen the manipulability ratings obtained through different instructions for the disagreements they generate. One limitation, however, is that these two summary statistics are not always available in the literature – especially for FM ratings. As pointed out in the introduction, pilot rating studies typically do not report item-level data to enable their further inspection; only summary statistics for different stimulus categories. Additionally, several rating studies do not provide SDs for their ratings (Amsel et al., 2012; Binder et al., 2016; Brodeur et al., 2010, 2012, 2014; Hoffman & Lambon Ralph, 2013; Medler et al., 2005; Ni et al., 2019; Wang et al., 2023). It is thus difficult to determine whether participants truly disagreed on midscale items. The issue can nevertheless be to a large extent circumvented if midscale manipulability ratings can be shown to predominantly result from disagreements. This would indeed allow us to detect disagreed-upon items solely based on their average ratings and would facilitate the interpretation of studies in which item-level data or SDs are not available.

The left panel of Figure 14 depicts the SDs plotted against the averages of all of the available unipolar manipulability ratings identified in the current work that provide both summary statistics (N = 23). For better readability, the averages are rescaled to [0, 1] and the SDs to [0, 0.5] based on their theoretical upper bound. The area in which the items generating the highest rate of disagreement are expected is highlighted in red for each study. The right panel of Figure 14 further presents the distribution of SDs for only midscale items in each dataset. The two plots indicate that some reliably midscale items ($SD < 0.25$) can be systematically found in SM ratings. Luckily, this dimension has been assessed through a relatively less diverse set of instructions compared to FM, and is predominantly reported by rating studies which provide both means and SDs. It is thus possible to directly analyse both summary statistics to detect the items generating disagreements. For FM ratings, however, virtually all items in the middle of the scales point to a disagreement in their underlying responses. The datasets presented in Figure 14 additionally cover a large variety of instructions used to assess this dimension. In the following section, midscale items for FM ratings

will thus be considered as resulting from a disagreement among raters when the relevant summary statistics are not available. Finally, the situation is a little less clear for MM ratings as only one dataset provides SDs. Based on this study and on the pattern of results observed for the two other types of instructions, we can nevertheless safely assume that a large portion of midscale items reflect a disagreement. We will additionally use comparisons with SM and FM ratings to gain a better understanding about what the MM dimension captures (see Section 4.3.3).

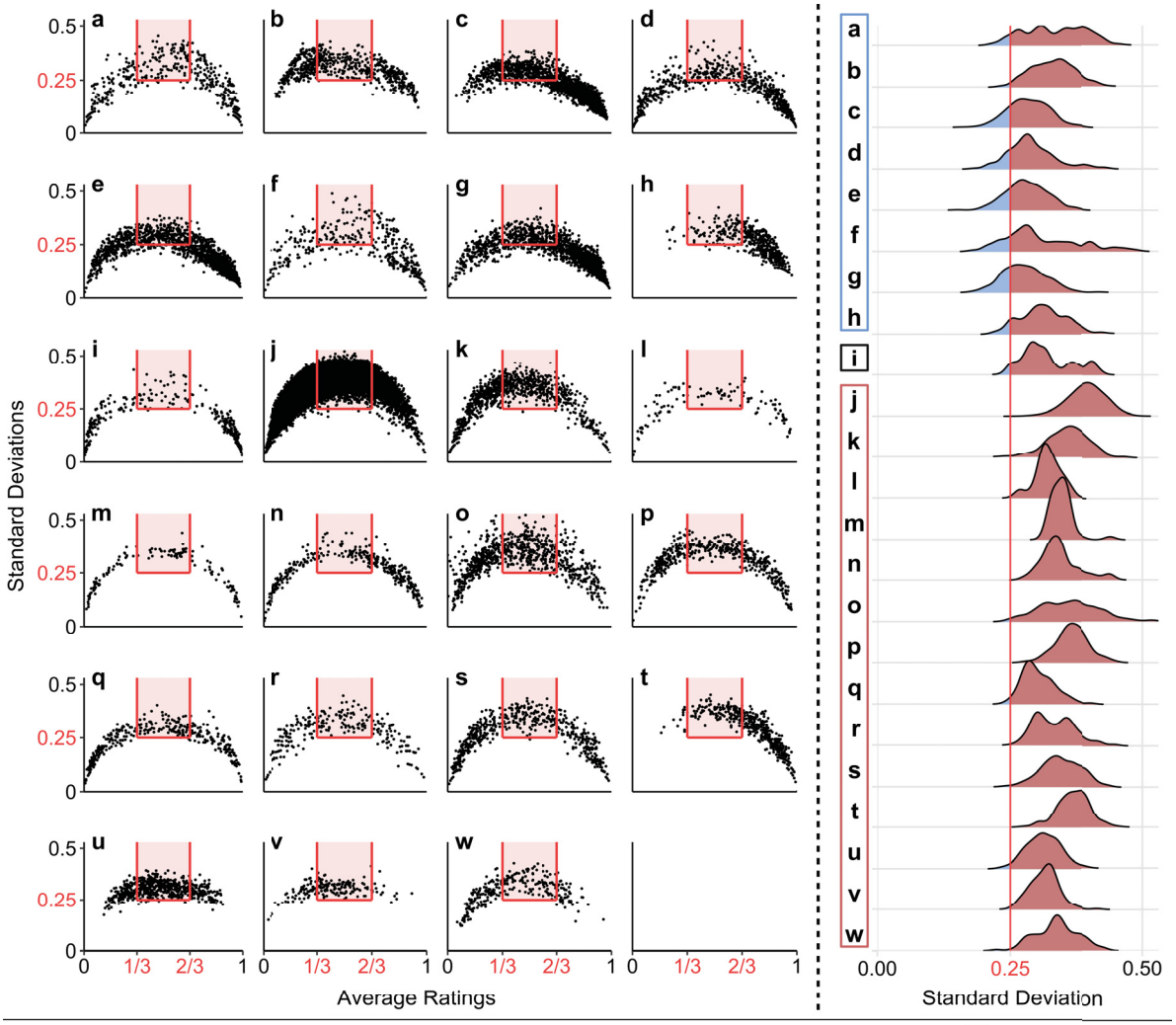
4.3 Dimensions of manipulability

This section offers a review of Likert-type rating scales used for the assessment of manipulability and a discussion of their validity in capturing its structural and functional components. To that end, the reviewed instructions are organised into three subsections relative to the dimension they most likely capture based on their wordings, namely structural (Section 4.3.1), functional (Section 4.3.2), and mixed (Section 4.3.3) manipulability. The latter subsection corresponds to instructions which explicitly refer to both structural and functional actions. As mentioned in Section 4.1, several studies do not explicitly assess the functional manipulability of objects, nor distinguish it from structural manipulations. However, their scales typically capture the extent to which objects are associated with actions – which, for manipulable objects, should be mostly functional. For the purposes of this review, these will thus be regarded as FM ratings. Within each subsection, instructions are further grouped based on the similarity of their wording to facilitate their analysis.

The approach used to evaluate the validity of the scales is largely based on the insights gained from the midscale disagreement problem (Section 4.2 and Chapter 3) and two general guiding principles. First, an instruction for a given dimension of manipulability should reliably yield high ratings for unequivocally manipulable objects, and low ratings for non-manipulable ones. Second, it should be as unambiguous as possible, and thus generate as little disagreements as possible overall. Each type of instruction is discussed with respect to potential ambiguities in its wording and to the disagreements found in the results it yields. We further highlight how different methodologies in data collection can affect the ratings. The analyses mainly rely on a graphical inspection of rating distributions for different semantic categories within each dataset. For instance, observing that a large number of animals have midscale average ratings when assessed through a given instruction will be interpreted as this category generating disagreements. This enables us to identify regularities in what leads to disagreements and to discuss their potential sources. The classification of items into semantic categories was achieved by aggregating and adjusting the category labels provided by eight different studies (Banks & Connell, 2022; Brodeur et al., 2010, 2014; Miklashevsky, 2018; Moreno-Martínez et al., 2014; Navarrete et al., 2019; Ni et al., 2019; Stoinski et al., 2023. The

Figure 14

Standard deviations against the mean ratings for all identified manipulability-related datasets (left) and ridge plots for the distribution of the standard deviations for midscale items (right)



Notes. The areas highlighted in red in the plots of the left panel correspond to values argued to be representative of a disagreement among raters, in which case the average ratings cannot be interpreted as the true position of the items on the scale. The ridge plot on the right shows the distribution of standard deviations for items falling within the middle third of the scale for each dataset. They are further organised according to their category of manipulability, i.e. structural (blue), mixed (black) and functional (red).

details are provided Appendix B). Eleven broad categories were selected for the current analyses. Tools, kitchen utensils and desk supplies were treated as a single category that should mainly include structurally and functionally manipulable objects. Furniture and appliances were similarly combined as they represent large household items that involve some degree of manual interaction. The other categories were musical instruments, weapons, wearable items (clothing), edible items (food), body parts, animals, vehicles, buildings or sections of buildings (building) and professions. As the analyses rely on extensive examinations of data and plots across a large number of datasets, their inclusion in the main text would be highly cumbersome. To maintain the clarity and focus of the narrative, the following sections provide a general assessment of the instructions while the detailed analyses are presented in Appendix B.

4.3.1 Structural manipulability ratings

Structural manipulability (SM) is a relatively recent dimension and has been assessed through fairly similar instructions (Table 5). It was introduced following the evidence for a dissociation between prehensile actions guided by bottom-up object characteristics (structural) and those related to an object's use (functional), and because previous manipulability scales mostly failed to capture the former (see Section 4.1). SM instructions thus aim to assess the extent to which the physical properties of an object (e.g. shape, size) allow for it to be grasped and/or picked up. As the instructions reviewed below also reveal, some specifics of what constitutes SM nevertheless remain unclear. For instance, some studies explicitly restrict structural actions to one hand (e.g. Amsel et al., 2012; Iachini et al., 2014), while others allow for both (e.g. Guérard et al., 2015; Rueschemeyer et al., 2010b). It is also not evident across studies whether SM necessarily implies that the object should be small and thus possible to grasp entirely and hold in one's hand(s), or if it is sufficient for it to have a graspable part. These uncertainties and variations in the phrasing of the instructions across studies lead to some degree of disagreement in the ratings. Nevertheless, the number of affected items remains relatively low compared to the FM instructions presented in Section 4.3.2.

4.3.1.1 Ease to grasp

SM has been most commonly assessed by asking participants to rate the ease with which they can grasp an object (Guérard et al., 2015; Heard et al., 2019; Iachini et al., 2014; Stoinski et al., 2023; Vingerhoets et al., 2009). Although these instructions appear to yield overall consistent ratings, a few results suggest that their interpretation differs across participants. For instance, some rather large objects such as pieces of furniture (e.g., *bookshelf*, *chest*, *table*. Guérard et al., 2015)

Table 5
Excerpts from the instructions used to assess structural manipulability

Category	Study	Instructions
Ease to grasp	Vingerhoets et al. (2009)	"healthy controls [...] were asked to and rate [each tool] [...] for graspability" (p. 482–483)
	Iachini et al. (2014)	"how easy is it to grasp the stimulus with one hand?" (p. 22)
	Guérard et al. (2015)^R	"Objects differ in the extent to which a person can grasp them using their hands. [...] The purpose of the present experiment is to rate objects regarding the ease with which a person can grasp them" (p. 456, in Appendix 2)
	Heard et al. (2019)^R	"The referents that words refer to vary in terms of how graspable they are. Something that is graspable can be grasped using one hand whereas something that is not graspable cannot. In this study your task is to rate each word's referent based on how graspable it is" (p. 10, in Appendix)
	Stoinski et al. (2023)^R	"How difficult/easy is it to grasp?" (p. 1589)
Likelihood to grasp	Amsel et al. (2012)^R	"How likely is someone to grasp this object with one hand?" (p. 1038, in Appendix A)
	Díez-Álamo et al. (2018)^R	"How probable is it that someone grabs or picks up this object with one hand?" (our translation of the original instructions in Spanish, p. 1641, in Appendix)
Ease to hold	Rueschemeyer et al. (2010b)	"participants were asked to rate words [...] with respect to [...] whether they could hold the object denoted by the word in their hands" (p. 1846)
	Stoinski et al. (2023)^R	"How difficult/easy is it to hold?" (p. 1589)
Ease to move	Guérard et al. (2015)^R	"The purpose of the present experiment is to rate objects regarding the ease with which a person can move them in space" (p. 457, in Appendix 2)
	Stoinski et al. (2023)^R	"How difficult/easy is it to move?" (p. 1589)
Ease to grasp for moving	Clarke & Lundington (2018)^R	"participants were asked to rate the ease with which they could grasp each object for the purpose of moving it" (p. 614)

^{R} Rating studies

and vehicles (e.g. *ambulance*, *motorcycle*, *wheelchair*. Heard et al., 2019) appear to cause disagreements in the judgements. Animals, on the other hand, are distributed across the entire range of the scale, with a large number of them nevertheless being in or close to its middle (e.g. *ant*, *monkey*, *snake*. Stoinski et al., 2023). It is important to note that the three datasets reporting ease to grasp ratings (Guérard et al., 2015; Heard et al., 2019; Stoinski et al., 2023) display some variability among themselves. This can be partly explained by their protocols, such as collecting ratings on a single or on multiple scales for a same item, or the nature of their stimuli (words, photographs or both). The three studies also differ in their instructions on one important aspect. Guérard et al. (2015) asked whether the objects could be grasped using the hands (thus leaving the possibility of using both open), while Heard et al. (2019) restricted it to one hand. Stoinski et al. (2023) do not specify this point. Assessing how these differences affected the ratings is unfortunately not entirely

possible with the available data.

These observations suggest that there are differences in how the *ease* or *ability* to perform the action is understood, with some participants responding rather generically (i.e. considering the possibility, in principle, of grasping an object), while others think about the items in real-life contexts (thus likely accounting for factors such as their animacy, dangerousness, or prevalence in the environment). Additionally, asking about the act of *grasping* alone can be ambiguous as it is difficult to judge whether it refers to the entire object or only to parts of it.

4.3.1.2 Likelihood to grasp

A variation of the previous instructions was proposed by Amsel et al. (2012; Díez-Álamo et al., 2018) who asked about the likelihood for someone to grasp an object with one hand. At first glance, these ratings appear very similar to those obtained with the 'ease to grasp' task. In fact, they correlate more with the 'ease to grasp' ratings than the three available 'ease to grasp' datasets do among themselves. The instructions additionally suggest that they might have the added benefit of capturing the extent to which objects lend themselves to being grasped during everyday interactions. Compared to the previous scale, this seems to partly resolve the disagreements observed for vehicles which are rated relatively lower, even though some can still be found in the middle of the scale. Animals are nevertheless once again distributed across the entire scale with a large number of them found in its middle or close (e.g. *bee*, *rabbit*, *spider*). Surprisingly, the current instructions also generate disagreements for a few seemingly graspable items such as some musical instruments (e.g. *banjo*, *violin*, *flute*) and weapons (e.g. *gun*, *rifle*, *shield*).

Although the instructions can be argued to be slightly closer to capturing SM on a conceptual level, the disagreements they generate suggest that they share similar ambiguities with the 'ease to grasp' task. Here as well, the *likelihood* to grasp an object might be interpreted both generically (i.e. the possibility of grasping something if someone tried) or in a realistic context (i.e. how likely someone is to grasp the object in everyday life). Additionally, the instructions used by the two rating studies specify that the action must be performed with one hand. This might lead to disparate judgements for objects which only have a graspable part, or those which are typically held with two hands during their use.

4.3.1.3 Ease to hold

In their pilot study, Rueschemeyer et al. (2010b) asked participants to rate whether they could hold the objects in their hands. This formulation is arguably less ambiguous than the previous as it implies that the object is small enough to be held. It would nevertheless likely not capture larger objects which have a graspable part. As the study does not provide item-level data, and because the

two stimulus categories were structurally manipulable objects, it is not possible to directly analyse the data these instructions yield.

The only rating study using a similar task is that of Stoinski et al. (2023) in which participants rated how difficult or easy it is to hold the objects. One important limitation of their instructions, however, is that they do not mention the hands (i.e. ease to hold *in the hands*). This can lead to ambiguity and to disagreements due to the diversity of meanings the verb holds, such as preventing an object from moving (e.g. *door*, *shelf*, *table*) or maintaining the hand on something (e.g. *touchpad*, *button*, *eye*). Without this clarification, objects which are not typically associated with SM can thus generate disagreements and end up towards the middle of the scale. It should also be mentioned that participants in Stoinski et al.'s (2023) study rated the same items on 11 different scales, three of which were related to SM (ease to grasp, ease to hold, and ease to move). The fact that these instructions are highly similar might have led participants to adjust their ratings and to respond differently than how they would have if given a single SM scale (as in, e.g. Díez-Álamo et al., 2018; Guérard et al., 2015; Heard et al., 2019). In line with this interpretation, we find indeed that the 'ease to hold' ratings provided by this study correlate more strongly with the other datasets than those they report for the 'ease to grasp'. Keeping these important points in mind, the distributions of different semantic categories on this scale are highly similar to those observed for the 'ease to grasp' ratings.

4.3.1.4 Ease to move

Guérard et al. (2015) argued that graspability ratings alone might not fully capture SM as "some objects difficult to grasp might nevertheless be easy to move by pulling or pushing (e.g. bicycle, wheelchair)" (p. 2). To address this, they have introduced ratings for the *ease to move* objects (see also Stoinski et al., 2023). Regardless of the conceptual relevance of this dimension for SM, its assessment on a separate scale seems to bring little additional information while introducing several confounds. Indeed, the ratings for the ease to grasp and the ease to move are highly correlated in both Guérard et al. (2015) and Stoinski et al. (2023), and no objects which can be considered to afford SM are rated as difficult to grasp but easy to move. Considering midscale items on Guérard et al.'s (2015) 'ease to grasp' scale, the examples cited by the authors (*bicycle*, *wheelchair*) are the only ones which can be argued to afford SM and rated high for their ease to be moved. Additionally, the instructions of neither of the two rating studies clearly state how the objects are moved (i.e. by hand) which likely explains the disagreements for several non-SM items generating disagreement or even receiving high ratings. Among these are notably some animals (e.g. *duck*, *fox*, *spider*) and, especially in Stoinski et al. (2023), vehicles (e.g. *ambulance*, *car*, *ship*). Overall, this dimension appears to be largely captured by other SM instructions and to miss

the critical aspect of being able to grasp or pick objects up – thus bringing little added value to the assessment of SM as a separate scale.

4.3.1.5 Ease to grasp for moving

The high correlation between the ease to grasp and the ease to move ratings observed by Guérard et al. (2015) led Clarke and Lundington (2018) to propose a combined version of the instructions. In their study, participants were asked to rate the ease with which objects can be grasped for the purpose of moving them. Conceptually, these instructions are at the crossroads between the other SM-related instructions, making them particularly interesting for this dimension's assessment. Indeed, they can be expected to resolve some ambiguity by specifying the purpose of the action (i.e. to move), and thus imply that the object can be held or that it has a graspable part. Unfortunately, Clarke and Lundington's (2018) data appear to contain several inconsistencies, making them difficult to analyse and to compare to other ratings. First and foremost, there is considerable disagreement among participants for a large number of clearly graspable and moveable items (e.g. *axe*, *feather*, *pineapple*, *scissors*, *syringe*). Second, the majority of items are found in the middle or in the high end of the scale. Only 7 out of the 480 of them have ratings on the low end, and display very high standard deviations. Given the patterns observed in the previously described scales, it is highly unlikely that these results are due to the instructions. One possibility – although rather unlikely as well – is that the Thai version of the instructions used for data collection contained ambiguities which are not apparent in the reported English translation. A more likely explanation is that the discrepancies stem from the items included in the study, the vast majority of which are manipulable objects. This can lead participants to adjust their ratings relative to the other items they have seen and to deviate from the instructions – especially if they are asked to use the entire scale for their judgments[32]. Although these instructions appear highly relevant for the assessment of SM, their validity thus remains to be examined.

[32] A slightly different example of this can be found in Villani et al. (2019). Among other variables, the authors collected body-object interaction ratings (Siakaluk et al., 2008a) for a set of 425 abstract nouns. As this scale is intended to capture the ease with which the human body can physically interact with objects, it would be expected to yield low ratings for abstract concepts. Surprisingly, however, most items are found towards the middle of the scale or on its high end (e.g. *pain*, *fear*, *energy*, *wish*). It thus appears that participants adjusted their judgments relative to the items they were presented instead of strictly following the instructions.

4.3.2 Functional manipulability ratings

Table 6

Excerpts from the instructions used for – or related to – the assessment of functional manipulability

Category	Study	Instructions
Action association	Gainotti et al. (2009) Gainotti et al. (2013)	"subjects were requested to evaluate [...] the relevance that, in constructing our knowledge of that item, had played a number of sensori-motor (visual, auditory, tactile, olfactory, and taste perceptions and motor activities) sources of knowledge" (p. 805, Gainotti et al., 2009)
	Rueschemeyer et al. (2010b)	"participants were asked to rate words [...] with respect to [...] whether they associated the object denoted by the word with an action" (p. 1844–1845)
	Proverbio et al. (2011)	"score whether the presented object evoked a motor association" (p. 2713)
	Carota et al. (2012)[a]	"participants had to rate the meaning of potential stimulus words with regard to a number of semantic variables [...] such as semantic relationship to action" (p. 1493–1494)
	Hoffman & Lambon Ralph (2013)[R]	"You will be asked how much you associate each item with a particular: [...] **Performed action** – movements you make when you interacting with the entity" (p. 1 of Appendix B, emphasis in original)
	Magri et al. (2021)[P]	"Participants [...] rated the degree to which the object made them think of moving their hands or other parts of their body" (p. 2)
Hand/arm action association	Carota et al. (2012)[a]	"participants had to rate the meaning of potential stimulus words with regard to a number of semantic variables [...] such as semantic relationship to action" (p. 1493–1494)
	Dreyer et al. (2015)	"Each word was rated for its semantic relatedness to hand/arm- [...] actions" (p. 5)
	Binder et al. (2016)[R]	"To what degree do you think of this thing as [...] being associated with [...] actions using the arm, hand, or fingers" (in online supplementary materials)
	Lynott et al. (2020)[R]	"To what extent do you experience WORD by performing an action with the [...] Hand/arm" (p. 1275)
	Magri et al. (2021)[P]	"Participants [...] rated the degree to which the object made them think of moving their hands" (p. 5)
	Repetto et al. (2023)[R]	"How much do you experience WORD through an action of [...] hand/arm" (p. 4037)

Continued on next page

Table 6

Excerpts from the instructions used for – or related to – the assessment of functional manipulability (Continued)

Category	Study	Instructions
Manipulability	Howard et al. (1995)	"items were rated for [...] manipulability" (p. 290)
	Kellenbach et al. (2003)	"rate 'how strongly you associate manipulation with each of the objects'" (p. 41)
	Medler et al. (2005)^{R} Fernandino et al. (2016)	"think about whether that word is associated with a manipulation or not. In this study, we define manipulation as a physical action done to an object by a person. [...] rate each word according to how relevant the concept of manipulation is to its meaning or purpose" (in online supplementary materials)
	Pecher et al. (2013)	"Objects differ in the extent to which people perform actions with them. Some objects always involve one or more specific actions, while with other objects, actions are rarely or never performed. The extent to which actions are carried out with an object is referred to as manipulability. The goal of the current task is to investigate the manipulability of a number of objects" (our translation of the original ratings in Dutch, personal communication, January 15, 2024)
	Desai et al. (2016)	"rate whether the noun refers to an object that can be physically manipulated" (p. 4051)
	Haddad et al. (2024)	"participants had to [...] rate [...] to what extent the object was manipulable" (p. 5)
Hand necessity for function	Tranel et al. (1997)	"Rate the extent to which use of human hands is necessary for this object to perform its function" (p. 1331)
	Moreno-Martínez & Adrados (2007)^{R}	"evaluate the degree of manipulation of each of these things. By degree of manipulation we mean the extent to which you believe that the hands are necessary for that thing to realise its function" (our translation of the original Spanish intructions, p. 4 of Appendix A)
	Moreno-Martínez et al. (2011)^{R} Moreno-Martínez & Montoro (2012)^{R} Moreno-Martínez et al. (2014)^{R} Navarrete et al. (2019)^{R}	"Participants were instructed to rate each item, assessing 'the degree to which using a human hand is necessary for this object to perform its function'" (p. 299, Moreno-Martínez et al., 2011)
	Miklashevsky (2018)^{R}	"estimate the words in the list by the necessity of use of human hands in order to make objects performing their typical functions" (p. 657, in Appendix 2)

Continued on next page

Table 6

Excerpts from the instructions used for – or related to – the assessment of functional manipulability (Continued)

Category	Study	Instructions
Ease to pantomime	Magnié et al. (2003)[R] Brodeur et al. (2010)[R] Brodeur et al. (2012)[R] Brodeur et al. (2014)[R, b]	"Could you easily mime the action usually associated with this object so that any person looking at you doing this action could decide which object goes with this action?" (p. 524, Magnié et al., 2003)
	Mahon et al. (2007)	"Suppose you were playing charades, such that one person had to identify an object/thing based on how another person mimed various actions that might be associated with that object/thing. You are asked to rate, for the following objects/things, how difficult it would be to play that game with these items" (p. 1 of supplementary materials)
	Campanella et al. (2010)	"Subjects were asked to judge how easy it was for them to mime the action commonly associated to the presented object so that anyone seeing that action could understand which object is associated to that action" (p. 1588)
	Wang et al. (2023)[R]	"rate the extent to which the meaning of a word can easily and quickly trigger corresponding body actions in your mind. Specifically, suppose you were playing a pantomime game in which one person had to identify a word based on how another person mimicked various actions that might be associated with its meaning" (p. 2)
	Guérard et al. (2015)[R] Heard et al. (2019)[R, c]	"Objects differ in the extent to which a person can think of an action involving that object. For some objects it is easier to think of an action than for others. The purpose of the present experiment is to rate objects regarding the ease with which a person can pantomime their use" (p. 457, in Appendix 2, Guérard et al., 2015)
	Clarke & Lundington (2018)[R]	"participants were asked to rate each item according to the ease with which they could mime its use when alone" (p. 614)
Grasp-use dissimilarity	Salmon et al. (2010)[R]	"rate the **extent to which the hand movements that you make to use the object differ from the hand movements that you make to pick it up**" (p. 85, emphasis in original)
	Lagacé et al. (2013)[R]	"[Participants] had to rate [. . .] the extent to which the posture of the hand to use the object differed from the posture of the hand to grasp it" (p. 775–776)

Continued on next page

Table 6

Excerpts from the instructions used for – or related to – the assessment of functional manipulability
(Continued)

Category	Study	Instructions
Number of actions	Lagacé et al. (2013)[R]	"participants were required to identify the number of actions that they could perform with each object [...]. An action was defined as a gesture that could be performed to use the object, with the action to grasp the object being excluded" (p. 776)
	Guérard et al. (2015)[R]	"determine the number of actions that you can perform with the object" (p. 457, in Appendix 2)
	Heard et al. (2019)[R]	"estimate the number of actions that you can typically perform with the word's referent. [...] although it may be possible to use the word's referent for other actions (e.g., using a screwdriver to hammer a nail) we ask that you rate the referent on the number of actions it would typically be used for" (p. 10, in Appendix)

[R] Rating studies providing item-level data.

[P] Studies with a pilot rating phase providing item-level data.

[a] The study reports ratings for action and arm relatedness but not the precise instructions used for the two dimensions.

[b] The instructions used by Brodeur et al. (2014) were slightly different: "... which object is associated with ..." (p. 3–4) instead of "... which object goes with ...".

[c] The instructions used by Heard et al. (2019) were slightly different: "... pantomime or act out the use of their referent" (p. 10, in Appendix) instead of "... pantomime their use" to

4.3.2.1 Action association

The most generic instructions among the FM variables require participants to rate the extent to which objects are associated to an action (see Table 6 for variations). They have been used both as an assessment of object manipulability (Proverbio et al., 2011; Rueschemeyer et al., 2010b) and of a concept's motor content more generally (Carota et al., 2012; Gainotti et al., 2009, 2013; Hoffman & Lambon Ralph, 2013; Magri et al., 2021).

An inspection of the studies which have collected action association ratings suggests that the instructions introduce considerable flexibility in their interpretation. For instance, objects labelled as FM in Rueschemeyer et al.'s (2010b) experiment had mostly midscale ratings. Some items identified as highly relevant to motor actions in Magri et al.'s (2021) pilot study were similarly found in the middle of the scale, while most were very close and thus likely displayed high disparity in the judgements as well. In contrast, Carota et al.'s (2012) summary statistics show that a large number of tools were rated on the low end of the scale along with foods and animals. To our knowledge, the only available dataset for action association ratings was provided by Hoffman and

Lambon Ralph (2013). In line with the previous observations, the majority of items are found either in or close to the middle of the scale. Critically, some clearly FM objects display disagreements (e.g. *clarinet*, *brush*, *machete*). The ranking of some of the items rated high also seems inconsistent with a reasonable assessment of FM (e.g. *escalator* and *train* have higher ratings than *axe*, *violin* and *typewriter*).

These instructions thus appear to yield unreliable ratings overall and suffer from important limitations in their wording to be regarded as a reliable assessment of FM. Indeed, they neither state that the actions to be considered should be specifically manual, nor that their goal is functional use. Importantly, several imply that the associated actions could be directly carried out by the entity being rated (e.g. animals, vehicles. Carota et al., 2012; Gainotti et al., 2009, 2013; Hoffman & Lambon Ralph, 2013; Proverbio et al., 2011; Rueschemeyer et al., 2010b). These elements can make the instructions highly ambiguous and likely sensitive to the specific context of each study (e.g. item-order effects, categories of objects rated by each participant).

4.3.2.2 Hand/arm action association

The 'hand/arm action association' dimension is typically found in the context of perceptual and motor strength norms (Carota et al., 2012; Binder et al., 2016; Dreyer et al., 2015; Lynott et al., 2020; Repetto et al., 2022), but have also been used by Magri et al. (2021) as an assessment of manipulability. As their name suggests, these instructions resolve one of the previous scale's limitations by specifying that only manual actions should be considered, although they lack a mention of the action's functional goal as well.

With the exception of Carota et al. (2012) and Magri et al. (2023), all of the studies in this category asked their participants to rate the items on other effector-specific action scales (e.g. foot/leg, face/mouth action associations). It is not entirely clear how this affects the judgements. Nevertheless, it can be reasonably expected that it led some participants to rate the prevalence of hand/arm actions relative to the other effectors instead of the extent to which they actually associate them with the items. A similar point can be made about the studies which have concurrently assessed the association to various perceptual modalities (e.g. Connell & Lynott, 2012). In all of these cases, participants might have responded differently if they were given a single scale.

A second issue concerns the nature of items included in the rating tasks. In one of the largest-scale studies to date, Lynott et al. (2020) normed words from several syntactic categories, including a large number of nouns and verbs (see also Repetto et al., 2023). The presence of action words (e.g. *pointing*, *touching*, *writing*, *applause*) appears to have heavily biased the ratings for nouns by generating disagreement among participants for FM objects. A large number of tools, utensils, supplies, instruments and weapons can indeed be found either in the middle (e.g.

flute, *harmonica*, *mug*, *sponge*, *syringe*, *toothbrush*) or close to the middle of the scale and with high standard deviations (e.g. *accordion*, *axe*, *joystick*, *key*, *scissors*). The dataset additionally displays disagreements for several other categories such as animals (e.g. *camel*, *otter*, *sheep*), furniture (e.g. *chair*, *bed*, *nightstand*), appliances (e.g. *television*, *refrigerator*, *washing machine*), vehicles (e.g. *boat*, *forklift*, *subway)* and buildings (e.g. *apartment*, *school*, *warehouse*). Similar results can be found in Repetto et al., 2023). In contrast, Binder et al.'s (2016) study contained relatively fewer verbs. Most importantly, their participants did not systematically rate action words as in the two other rating studies and their judgements were thus less likely to be affected by their presence. In line with these observations, Binder et al. (2016) report generally higher ratings for tools, utensils and instruments, as well as lower ratings for animals and buildings. A few categories nevertheless appear to continue to generate disagreements, notably foods (e.g. *banana*, *bread*, *chocolate*), profession names (e.g. *actor*, *doctor*, *farmer*), and some vehicles (e.g. *bicycle*, *car*, *elevator*).

The methodological issues outlined above constrain the conclusions that can be drawn about these instructions and show that their interpretation can be highly contextual. Nonetheless, Binder et al.'s (2016) results suggest that they can yield mostly valid ratings under adequate task conditions, even though they present some inherent ambiguity. The disagreements observed in this study would likely be partly resolved by specifying the action's goal – i.e. functional use.

4.3.2.3 Manipulability

The studies grouped into this category have taken a more direct approach by assessing whether objects are manipulable (Desai et al., 2016; Haddad et al., 2024; Howard et al., 1995) or the extent to which they are associated to manipulation (Fernandino et al., 2016; Kellenbach et al., 2003; Medler et al., 2005; Pecher et al., 2013). *Manipulation* is here generally used implicitly in the sense of manual actions performed with objects (but see below for Fernandino et al., 2016; Medler et al., 2005). These instructions are thus conceptually very similar to those assessing the 'hand/arm action association'.

Ratings with manipulability instructions have been predominantly collected in pilot rating studies which do not provide item-level data, thus limiting the possibility of their analyses. The available data point to some degree of validity for these instructions as manipulable objects appear to be mostly rated on the high end of the scale – although some also generate disagreement. The results are nevertheless less clear for other types of items. For instance, Howard et al. (1995) provide summary statistics for their stimuli when grouped either as high and low in operativity[33],

[33]Operative items are defined as "discrete and separate from the surrounding context, easy to manipulate, firm to the touch, easily available to several sense modalities" (p. 214, Gardner, 1973).

or as animate and inanimate. The majority of low operativity items seem to have received midscale ratings. A large number of both animate and inanimate entities similarly appear to have caused a disagreement in the judgements. Desai et al. (2016) report two sets of stimuli used for correlational analyses with fMRI data. The majority of them seem to have midscale ratings and thus to have been disagreed upon as to whether they can be physically manipulated.

Pecher's (2013) and Medler et al.'s (2005. See also Fernandino et al., 2016) studies distinguish themselves by having further defined the concept of manipulability in their instructions. In the former case, manipulability was explicitly defined as the "extent to which actions are carried out with an object" (personal communication, January 15, 2024). These instructions are thus particularly similar to the 'hand/arm action association' scale. Although the study's two stimulus lists did not contain any midscale items, those labelled as high in manipulability had ratings close to the middle of the scale and thus likely generated some degree of disagreement. To our knowledge, the only rating study in the present category providing ratings was conducted by Medler et al. (2005) who defined "manipulation as a physical action done to an object by a person" (online supplementary materials. The same instructions were later used in the pilot study of Fernandino et al., 2016). This choice is curious as manipulability is typically used in reference to actions performed *with* objects, not *on* them. A closer inspection of their ratings indeed suggests that this formulation introduced considerable ambiguity to the task. Almost all rated items are found either on the low end of the scale or in its middle. Most importantly, a large number of clearly FM objects – among others – appear to have been disagreed on (e.g. *axe*, *fork*, *hammer*, *knife*, *pliers*, *spatula*).

As with the previous two types of reviewed scales, the lack of data and methodological differences across studies do not allow to clearly determine the strengths and limitations of the current manipulability instructions. It is nevertheless possible to note that the nature and goal of the assessed manipulations are once again ambiguous, as it is not clear if they refer to the use of objects or more generally to the possibility of handling them. Moreover, the manner in which the term "manipulation" is employed during everyday language does not necessarily imply either functional actions (e.g. simply handling an object), nor manual actions more generally as it can signify the control, management or utilisation of something (e.g. to manipulate information, data, people). This likely represents an additional source of ambiguity for some items and can thus lead to disagreements.

4.3.2.4 Hand necessity for function

The 'hand necessity for function' instructions were introduced by Tranel et al. (1997) and popularised by Moreno-Martínez and collaborators in a series of rating studies (Moreno-Martínez & Adrados, 2007; Moreno-Martínez et al., 2011, 2014; Moreno-Martínez & Montoro, 2012). They

have also been used more recently by Miklashevsky's (2018) and Navarrete et al.'s (2019) rating studies. As their name suggests, these instructions assess the "extent to which use of human hands is necessary for [objects] to perform [their] function" (p. 1331, Tranel et al., 1997). They thus make an explicit reference to manual actions and to their goal (functional), resolving the main ambiguities highlighted for the previous instructions.

The data provided by rating studies strongly suggest that these instructions reliably capture FM objects such as tools, utensils, instruments and weapons. Some categories nevertheless display disagreements for a significant number of items. In Miklashevsky (2018), for instance, wearable items (e.g. *headphones*, *necklace*, *slipper*, *helmet*) and foods (e.g. *aubergine*, *lemon*, *mushroom*, *raisin*) are generally found in the middle of the scale or close to it on the high end of the scale. Vehicles are more generally distributed across the middle (e.g. *canoe*, *helicopter*, *locomotive*, *tram*) and high (e.g. *airplane*, *bus*, *motorcycle*, *ship*) ends of the scale. Finally, some buildings appear to also generate disagreements across datasets (e.g., in Moreno-Martínez et al., 2014, *church*, *factory*, *mill*, *skyscraper*).

The observed disagreements can likely be explained partly by the inherent ambiguity in what constitutes an item's function. Indeed, objects that have clearly identified functions (e.g. tools, instruments) appear to be reliably rated high on the scale. Other items (e.g. clothing, food), however, have less well-defined functions and can result in disagreements in the judgements. An additional limitation of the present instructions is that they do not specifically capture FM objects on the high end of the scale. Among the discussed examples, vehicles typically require operators – and the use of their hands, while buildings are generally associated with activities that can involve the hands. Such items can thus receive midscale or even high ratings, without necessarily being strongly associated with functional actions for participants – at least in the sense of object manipulation. Although these instructions appear conceptually closer to the assessment of FM than the previous ones, their phrasing thus appears to generate ambiguity for some items as well.

4.3.2.5 Ease to pantomime

The 'ease to pantomime' instructions are conceptually different from those previously discussed in that they were originally introduced as an assessment of the extent to which the actions associated with objects are distinctive. They were proposed by Magnié et al. (2003) and have since been one of the most commonly used operationalisations of manipulability in the norming literature (Brodeur et al., 2010, 2012, 2014; Campanella et al., 2010; Mahon et al., 2007; Wang et al., 2023). These instructions require participants to judge how easily they could "mime the action usually associated with [an] object so that any person looking at [them] doing this action could decide which object goes with this action" (p. 524, Magnié et al., 2003. For slightly different

wordings, see Mahon et al., 2007; Wang et al., 2023). More recently, Salmon et al. (2010) noted that this definition results in some manipulable objects to be rated low on the scale due to their associated actions not being distinctive (notably edible items). This observation led Guérard et al. (2015. See also Clarke & Lundington, 2018; Heard et al., 2019) to propose a modified version of the instructions so as to more accurately capture FM. Their task did not mention how recognisable the action would be if mimed and instead focused on the ease with which "a person could think of an action involving that object" (p. 457, in Appendix 2. See also the instructions used by Wang et al., 2023).

Rating studies which have used the original instructions (i.e. recognisable pantomime. Brodeur et al., 2010, 2014; Magnié et al., 2003) show some differences in their data for FM objects[34]. Indeed, tools, utensils, instruments and weapons are generally found on the high end of the scale in Magnié et al. (2003), whereas they have mostly midscale ratings in Brodeur et al. (2010, 2014. E.g. *brush*, *lighter*, *pliers*, *frying pan*, *knife*, *guitar*, *sword*). Body parts, on the other hand, are mostly found midscale in Magnié et al. (2003) but on the high end in Brodeur et al. (2014. E.g. *ear*, *eye*, *lips*, *nose*). A number of other categories appear to cause disagreements in all datasets. In Brodeur et al., (2014), for instance, several clothing items (e.g. *hat*, *shirt*, *shoe*), vehicles (e.g. *airplane*, *car*, *sailboat*), furniture (e.g. *bookshelf*, *chair*, *mattress*) and appliances (e.g. *fridge*, *stove*, *washing machine*) can be found in the middle of the scale.

The three studies providing ratings for the modified version of the instructions (Clarke & Lundington, 2018; Guérard et al., 2015; Heard et al., 2019) also show significant variability among themselves. Clarke & Lundington's (2018) and Heard et al.'s (2019) results nevertheless suggest that the differences stem from more general issues with these two datasets, while the ratings collected by Guérard et al. (2015) appear relatively more reliable (for a discussion, see Section 4.3.1.5 and Appendix B). This latter study's data show that the modified instructions overall result in FM objects (tools, utensils, instruments, weapons) to be more consistently rated high on the scale. Once again, however, some body parts (e.g. *chest*, *eye*, *lips*), vehicles (e.g. *car*, *plane*, *truck*) and furniture (e.g. *bench*, *bookshelf*, *dresser*) are disagreed upon. In contrast to the original instructions, edible items (e.g. *cake*, *orange*, *strawberry*) are also generally found in the middle of the scale.

Similar to most previous instructions, an important limitation of both the original and the modified versions of the pantomime task is that they fail to specifically target manual functional actions. Guérard et al. (2015; Heard et al., 2019), for instance, asked their participants to rate how easy it would be to pantomime the use of objects. They thus did not exclude uses which do not involve the hands – as is also evident from one of the two examples the authors used as anchors

[34]The discrepancies might originate from the relatively more specific scale labels used in Magnié et al.'s (2003) study and to how the ratings were computed. Indeed, the middle option corresponded to "I do not know" and was omitted when averaging the responses for each item.

for the high end of the scale (*chair*, because it is possible to sit on it). The original instructions more generally mention actions which are typically associated with objects. In addition to opening the possibility for non-functional and non-manual actions, this formulation might also generate disagreements in the ratings for items which are not associated with clearly identified actions. Moreover, the framing of the task as a common pantomime game raises concerns about how the participants responded. Indeed, miming an object can involve iconic gestures which would not be typically performed during real-life interactions (e.g. extending the arms to represent an airplane, mimicking scissors with two fingers). Some participants might have thus rated the items based on how well they could make others guess them in a real pantomime game, without necessarily considering the actual actions they would perform with them (as it is notably apparent in the ratings of body parts). The modified task used by Guérard et al. (2015; Heard et al., 2019. See also Wang et al., 2023) is particularly problematic in this respect because it starts with a focus on the ease with which participants can think of an action involving the object, followed by instructions to rate how easily its use can be pantomimed. This can likely lead to different interpretations, such as assessing how easily the object evokes either actual or iconic actions, but also how easily the action could be recognised in a pantomime context.

Overall, the two versions of the 'ease to pantomime' instructions present limitations regarding both their wording and the data they generate to be regarded as a reliable assessment of FM. As Magnié et al. (2003) initially proposed, the distinctiveness of the functional manipulations can nevertheless be argued to be an important aspect of FM. Given adequate adjustments to the original instructions – particularly to specify the nature of the actions to be considered, this dimension can thus be relevant for assessing the distinctiveness of the functional actions associated with an object.

4.3.2.6 Grasp-Use dissimilarity

Salmon et al. (2010) introduced the 'grasp-use dissimilarity' instructions in order to specifically capture FM. In their task, participants were first asked to rate the ease with which they could grasp and use a number of objects on a [1,5] scale (see Section 4.3.3). For items rated with a 3 or above, they were then presented with a second scale assessing "the **extent to which the hand movements that** [they] **make to use the object differ from the hand movements that** [they] **make to pick it up**" (p. 85, emphasis in original). Similar instructions were used by Lagacé et al. (2013) in a later rating study, although their task assessed the difference between hand postures instead of hand movements. Interestingly, the two studies differ in their theoretical grounds for using these instructions. In contrast to Salmon et al. (2010) who used them to capture FM, Lagacé et al. (2013) justified their choice based on studies suggesting that SM and FM compete during

object recognition (e.g. Bub et al., 2008; Jax & Buxbaum, 2010). Their goal was thus more directly to provide ratings that allow the detection of objects with different grips for the two types of manipulations.

An inspection of Salmon et al.'s (2010) dataset reveals that very few items have a high rating on the scale, while the majority are found in the middle. Almost all categories, and especially FM items such as tools (e.g. *drill*, *scissors*, *tape measure*), desk supplies (e.g. *binder*, *pen*, *ruler*), instruments (e.g. *harmonica*, *maracas*, *tambourine*) and weapons (e.g. *axe*, *bow*, *gun*) appear to generate disagreements. A specific issue with this study is that participants only rated the objects that they judged as graspable and usable on the first scale. As a result, several objects that are not graspable but could be expected to be associated to FM did not receive ratings for 'grasp-use dissimilarity (e.g. *cash register*, *harp*, *photocopier*, *piano*, *roulette wheel*). Additionally, the authors' method appears to have resulted in several graspable but not FM objects to be rated on the scale (e.g. *ant*, *fly*, *leaf*, *snail*). Few items were also rated high in Lagacé et al.'s (2013) study. Compared to Salmon et al. (2010), however, relatively more items are consistently rated on the low end of the scale. Critically, few FM objects appear to have generated disagreements and are generally found on the low end of the scale. Some other categories are nevertheless found midscale, such as clothing (e.g. *bag*, *hat*, *necklace*) and some edible items (e.g. *egg*, *grapes*, *toast*). It should be noted that this dataset contains very few, if any, exemplars for several of the inspected categories (i.e. instruments, body parts, animals, vehicles, furniture and appliances, buildings, professions). It is thus not possible to determine whether these items would generate disagreements.

An important limitation of these instructions is that they do not clearly differentiate SM from FM in their wording. For FM, both rating studies state that the action's goal is to use the object. For SM, however, they simply mention grasping the object or picking it up, without any reference to an end-goal. The possibility that the latter action is intended for the future use of the object is thus left open. As most FM objects also require to be grasped or picked up during everyday interactions, the instructions can be easily interpreted as targeting different phases of the movements necessary to use an object, instead of two different types of manipulation. This issue can be likely circumvented by specifying a goal for SM such as moving the object, which has been the main definition used by experimental studies investigating the interactions between SM and FM during object processing (e.g. Jax & Buxbaum, 2010; Valyear et al., 2011; Wamain et al., 2018). As these studies also suggest, 'grasp-use dissimilarity' ratings can be a valuable dimension to be normed. They nevertheless do not essentially capture FM as such and should be treated as complementary.

4.3.2.7 Number of actions

This dimension was introduced by Lagacé et al. (2013) to assess the number of actions that could be performed to use objects, and to investigate if it is related to how much participants agree about the specific grip shape necessary to use them. Guérard et al. (2015) also collected 'number of actions' ratings, although to specifically capture FM in conjunction with 'ease to pantomime' ratings. Finally, Heard et al. (2019) were interested in this variable as a potential predictor of the more generic body-object interaction ratings (see Section 2.4 and Chapter 3. Siakaluk et al., 2008a).

The 'number of actions' instructions do not lend themselves to an analysis of the disagreements they generate as they use a numerical Likert-type scale instead of a unipolar one. A number of observations can nevertheless be made about the wording of the instructions. First, none of the studies specifically mention manual actions in their task and Guérard et al.'s (2015) additionally omits their functional goal. It is thus likely that these ratings capture actions performed with other effectors as well. A second potential issue with Guérard et al.'s (2015) and Lagacé et al.'s (2013) studies is that the framing of their instructions can lead participants to think of actions that they would not otherwise perform with the objects during everyday interactions. In contrast, Heard et al. (2019) asked them to only rate the items on "the number of actions [they] would typically be used for" (p. 10, in Appendix). This added precision can be expected to both increase the data's overall validity, and to reduce the variability across participants that could arise from different interpretations of what counts as an action.

Two objects might receive comparably high ratings for the extent to which they are associated with manual actions for their use, but differ in the number of actions that can be performed with them (for a similar point, see Guérard et al., 2015). The 'number of actions' ratings can thus be seen as an assessment of the diversity of FM associated with an object and can prove valuable if used in conjunction with other scales assessing FM.

4.3.3 Mixed manipulability ratings

The studies discussed in this Section have used instructions that refer to both structural and functional manipulations (Table 7). This approach was introduced by Salmon et al. (2010) who argued for the importance of accounting for object graspability in the assessment of manipulability. Their task required participants to "**rate the manipulability of the object according to how easy it is to grasp and use the object with one hand**" (p. 85, emphasis in original) and several other studies have used similar instructions since (Bocanegra et al., 2017; Federico et al., 2023; Godard et al., 2019; Righi et al., 2014; Suárez-García et al., 2021; Tettamanti et al., 2017). At face value, the

wording of these instructions suggests that they assess the extent to which objects are both FM and SM. In line with this, most experiments have used them to capture the manipulability of objects, i.e. their association to FM (Bocanegra et al., 2017; Godard et al., 2019; Ni et al., 2019; Righi et al., 2014; Suárez-García et al., 2021). Curiously, however, Salmon et al. (2010) and other studies (Federico et al., 2023; Tettamanti et al., 2017) have used this scale for SM only.

Table 7
Excerpts from the instructions combining structural and functional manipulations

Study	Instructions
Salmon et al. (2010)^R	"**rate the manipulability of the object according to how easy it is to grasp and use the object with one hand**" (p. 85, emphasis in original)
Bocanegra et al. (2017)	"participants were asked to rate the extent to which each depicted item could be grasped and employed in a manual action" (p. 38)
Ni et al. (2019)^R	"Object manipulability was defined as 'the degree to which you are able to grasp or functionally use the item with hands in your daily life'" (p. 5)
Godard et al. (2019)	"participants [. . .] rated [. . .] 'the manipulability of the object according to how easy it is to grasp and use the object with one hand'" (p. 2802)
Federico et al. (2023)	"We asked the jury to rate each object from the study's stimuli on 2 Likert scales ranging from 1 (extremely difficult to grasp and use, extremely unrealistic) to 5 (extremely easy to grasp and use, extremely realistic)" (p. 8)
Righi et al. (2014)	"subjects were instructed to judge each stimuli [. . .] for [. . .] Manipulability that was defined as 'the extent to which the objects are capable of being grasped and manipulated by one hand'" (p. 242–243)
Tettamanti et al. (2017)	"Indicate the easiness with which is possible to grasp, pick up and manipulate the entity presented in the image" (personal communication, April 26, 2024)
Suárez-García et al. (2021)	"participants were asked to rate how graspable and manipulable each item was" (p. 4)

^{R} Rating studies

To our knowledge, only Salmon et al. (2010) and Ni et al. (2019) have provided ratings for these instructions. The latter study has used a slightly different formulation than Salmon et al. (2010) by asking participants to rate their ability to grasp *or* use objects, instead of their ease to grasp *and* use them. Logically, Salmon et al.'s (2010) instructions should thus capture items which are strictly both FM and SM, while objects which are either SM or FM could be rated high in Ni et al. (2019). An inspection of the two datasets nevertheless suggests that Salmon et al.'s (2010) ratings reflect SM only, while Ni et al.'s (2019) are closer to an assessment of FM, but not SM. For instance, graspable edible items (e.g. *apple*, *carrot*, *tomato*) are rated high in Salmon et al. (2010) but generate disagreements in Ni et al. (2019). Animals (e.g. *duck*, *spider*, *turtle*), on the other hand, are rated low in Ni et al. (2019) but are either found in the middle or on the high end of the scale in Salmon et al. (2010) – as with most other SM ratings discussed in Section 4.3.1. Correlational analyses also reveal that Salmon et al.'s (2010) ratings are highly correlated with SM

ratings but to a lesser extent with FM ones, while the inverse pattern is observed for Ni et al.'s (2019) data.

These results strongly suggest that mixed manipulability ratings can result in counterintuitive and unpredictable results that might not reliably capture the intended dimension of manipulability. They are also likely sensitive to the overall context of studies and to how the rating task is presented, which could cue participants to weigh one type of manipulability more than the other in their judgements. As such methodological details and item-level data are typically not reported, it is sometimes impossible to determine the dimension ultimately being assessed by different studies using these instructions. Overall, it thus appears more appropriate to assess manipulability through distinct SM and FM scales.

4.4 Discussion

The goal of the present work was to provide a review and analysis of the most common Likert-type rating task instructions used for the assessment of object manipulability. In light of a distinction between structural and functional manipulations (respectively SM and FM. Section 4.1), different types of instructions were discussed regarding their validity in capturing one of these two dimensions (Section 4.3). This was mainly achieved by directly examining the average ratings provided by available datasets, and particularly the items found towards the middle of the scale. Indeed, we have argued that a rating in the middle third portion of unipolar Likert-type scales typically results from a significant disagreement among participants (Section 4.2). Investigating the distribution of ratings for different semantic categories enabled us to detect regularities in the items that consistently received high or low ratings, and those that generated disagreements. It was thus possible to assess the ability of each type of instruction to correctly capture manipulable and non-manipulable items, as well as to further identify potential ambiguities in their wording.

A few studies have previously suggested the diversity in how manipulability has been operationalised, but its extent has been largely unknown (e.g. Guérard et al., 2015; Salmon et al., 2010). The current review highlights a striking heterogeneity across the literature and sometimes important differences in the data that different instructions yield – especially for FM. We have identified 9 studies which have assessed SM through 5 different instructions and 36 studies assessing FM with broadly 7 types of definitions. Eight further studies have referred to both SM and FM in their operational definitions, but disagreed as to which dimension their ratings capture. It is important to point out that our literature review revealed 7 additional instructions for FM that were not discussed (Godard et al., 2019; Iachini et al., 2014; Mahon et al., 2007; Righi et al., 2014; Tettamanti et al.,

2017; Tranel et al., 1997; Wolk et al., 2005). These were generally similar to the reviewed instructions but their wording also differed from them in important aspects. To our knowledge, these have only been used once in pilot rating studies and thus do not lend themselves to the analyses performed in the current work.

Most instructions assessing SM were overall very similar in both their wording and the data they generated. There nevertheless appears to be a general uncertainty among studies as to whether SM refers to prehensile actions performed with only one hand or both. Some of the observed results also suggest that relatively big objects which can have a graspable part (e.g. furniture, appliances) tend to generate disagreements among participants. Another consistent observation across studies is that raters appear to disagree in their judgements for animals. We have argued that the 'ease to grasp for moving' instructions introduced by Clarke & Lundington (2018) are conceptually the most appropriate for this dimension as they should partly resolve some of the inherent ambiguities with the wording of the other definitions and because they combine the two components of SM proposed by Guérard et al. (2015) into a single scale. Additionally, they are the only instructions specifying the goal of the action (i.e. to move), thus clearly distinguishing them from other types of object-directed interactions. It is also highly likely that they would resolve the disagreements for animals. Unfortunately, however, these suggestions remain speculative as the data provided by Clarke & Lundington (2018) display a number of validity issues which do not seem to be caused by the instructions.

The lack of a consensus in manipulability ratings was much more apparent in FM instructions. Among these, several are conceptually similar and broadly assess the extent to which objects are associated with actions – although they display considerable differences in their ratings due to their wording (action association, hand/arm action association, hand necessity for function, and manipulability). Others have radically different framings and seem generally inadequate to capture FM on their own (ease to pantomime, grasp/use dissimilarity, and number of actions). However, we have argued that these latter ratings can offer valuable information on FM that cannot be obtained through a single scale, namely the distinctiveness and diversity of the actions (respectively through the ease to pantomime and the number of actions ratings), as well as their competition with SM during object processing (grasp-use dissimilarity ratings). As Guérard et al. (2015) have also suggested, it thus appears that FM is best captured by considering these different ratings as complementary. This approach nevertheless requires caution as these variables can be expected to be strongly correlated, and should thus not be simultaneously included as covariates in regression models without proper diagnostic measures[35].

[35]The inclusion of highly correlated variables is a well-known problem in regression and can result in strongly biased estimates (Friedman & Wall, 2005; Hsu & Chiang, 2020; Nickels et al., 2022). A good example of this comes directly from Guérard et al.'s (2015) study. The authors fit a multiple regression model on naming latencies using two highly

One surprising, yet recurring observation with FM instructions was that a large number of studies failed to mention that the actions being assessed were manual, for functional use, or both. As we have argued, this can lead to disagreements or inconsistent ratings for a large number of items, as it allows for other actions to be considered and can result in different interpretations of the task. A more general issue with FM is that a rating task which aims to capture this dimension needs to specifically target the use of objects. As several reviewed instructions have revealed, objects which have clearly defined uses (e.g. tools) typically receive high ratings. However, what 'the use' refers to for some items (e.g. clothes, food) is much more ambiguous and can result in disagreements. There is thus an inherent limitation to the validity of FM ratings which appears difficult to resolve.

The final category of instructions discussed in the current work jointly referred to both SM and FM. Conceptually, these ratings can be regarded as capturing only graspable FM objects. However, their use has been inconsistent across studies, with some considering them as an assessment of SM, while others have taken them to represent FM. Our analysis additionally revealed that the results obtained with these instructions are largely counterintuitive and that it is difficult to predict which dimension they truly capture. We have thus argued that separate SM and FM scales should be preferred to such mixed instructions as they are generally more reliable.

Overall, our results clearly illustrate the significant disparity in manipulability's operationalisation, and raise concerns about the validity and comparability of studies investigating this variable through different instructions. Although the current work focused on single-item manipulability ratings, its results provide valuable insights that go beyond this scope. First, our analyses based on the midscale disagreement provide a direct method to investigate the validity of Likert-type norms and can be extended to any such variable. Regarding the study of manipulability, the ambiguities identified in the reviewed instructions can additionally prove useful for a more informed reading of related scales that use a similar vocabulary, such as manipulation familiarity/experience ratings (e.g. Kithu et al., 2021; Mahon et al., 2007; Ni et al., 2019; Rio, 2021; Vingerhoets et al., 2009) and various manipulation similarity judgements (e.g. Almeida et al., 2023; Helbig et al., 2006; Kalénine et al., 2012; Ruotolo et al., 2020; Kithu et al., 2021; McNair & Harris, 2012). Finally, the present review and discussions refine the assessment of SM and FM by drawing attention to several key dimensions underlying these variables and that are susceptible to play a role in object processing – especially regarding FM (distinctiveness, diversity). We hope that this work will serve as a common ground for researchers investigating the role of manipulability and will foster further discussions on methodological practices.

correlated SM variables ($r(557) = .973, p < .0001$) and several other predictors. The two variables surprisingly had significant but opposite relationships with naming latencies in the model.

The detailed examination of a large number of studies for the current review led us to a number of concerning observations regarding their methodologies. As suggested in the introduction, studies justifying their operationalisation of manipulability are the exception rather than the norm, which is particularly surprising given the exceptionally high number of different instructions found for this variable. Additionally, most studies fail to give a detailed description of their rating tasks, especially those with a pilot rating phase. In a few cases in which supplementary materials were available, we have sometimes observed that the rating instructions were not faithfully described in the main text of the article. We have also detected misattributions that could lead to confusion, with the instructions attributed to studies that did not use the same definition. Taken together, these issues make reliably tracing the methodologies of different studies highly challenging and leave the impression of a general neglect of rating-based stimulus selection procedures despite their crucial role for experimental validity.

During our analysis of the rating datasets, we have also identified broader methodological issues with Likert-type norms. First, participants do not appear to strictly follow the instructions, rather to adjust their responses relative to the sample of items they have been asked to rate (see Section 4.3.1.5 and footnote 31). It is thus paramount to provide participants with good anchors, as well as a randomised and balanced set of stimuli which can be expected to be distributed across the scale. A second issue concerns the lexical category of the items included in studies collecting ratings for words. In the case of manipulability, most rating studies have only used nouns. Some, however, have included other parts of speech (e.g. verbs, adjectives) which seems to have significantly affected participants' judgements for nouns (see Section 4.3.2.2). This suggests that it is more appropriate to collect ratings for different lexical categories separately, or to justify their joint inclusion otherwise. Finally, asking participants to rate the same item on multiple related scales can lead them to focus on small differences between instructions and to add ambiguity to what is being assessed (see Sections 4.3.1.3 and 4.3.2.2). Experimental fatigue is also rarely considered, with some studies having either a large number of scales (for an extreme example, participants judged each item on 65 different dimensions in Binder et al., 2016) or items, which might lead to inconsistent ratings that would likely not be detected by common reliability analyses. Indeed, to our knowledge there exists no systematic study of how reliability metrics used by rating studies behave under different response patterns – notably disagreements. There are thus no references or agreed upon standards to detect quality issues with subjective norms. It is important to stress that the observations discussed here are only preliminary and result from indirect analyses. They should thus be taken with caution until they are examined more thoroughly. Given the ever-increasing number and size of rating studies, the resources dedicated to them, and their importance for experimental validity, these remarks nevertheless show that there is a pressing need for methodological research

to establish clear guidelines and to avoid unusable datasets – as well as unreliable experimental results.

4.5 Conclusion

Our review of how manipulability has been defined and assessed highlights considerable differences across the literature, as much in their instructions as in the data that they generate. Functional manipulability ratings in general also appear to cause significant disagreements in the ratings. How these issues affect the experiments using manipulable objects as stimuli is nevertheless not entirely clear – and quite more difficult to assess than more established and systematically studied variables (e.g. body-object interaction, Chapter 3). Indeed, some experiments have collected manipulability ratings for stimulus validation. This implies that the stimuli were preselected by the researchers based on a more or less intuitive definition of manipulability. The validity of the ratings in such cases thus does not necessarily affect the reliability of the experimental results. On the other hand, studies that entirely relied on manipulability ratings for their stimulus selection or analyses remain subject to validity concerns depending on the instructions that were used and on the extent to which they generated disagreements.

Although our observations are preliminary, the current work is also highly informative about the methodological practices of norming studies and how they might affect their results. We have notably argued that the items included in the rating tasks could influence participants' judgements. In our conclusion of Chapter 3, we had pointed out that the comparison between Pexman et al.'s (2019) body-object interaction ratings and the combined datasets of Bennett et al. (2011) and Tillotson et al. (2008) displayed a systematic relationship, with items being generally rated higher in the former. In light of the present analyses, it is likely that this divergence resulted from the inclusion of different parts of speech (i.e. adjectives, adverbs, nouns, and verbs) in Pexman et al.'s (2019) study. Indeed, none of the syntactic categories except for nouns represent entities with which it is possible to interact with. It is possible that participants rated nouns relative to these categories, which would bias their judgements to be relatively higher on the scale. We should nevertheless stress that this does not necessarily invalidate our arguments about the measurement error for Likert-type ratings as it introduces a confounding source of variability; it only shows that it is difficult to gain insights about the issue by comparing rating studies with different methodologies.

Chapter 5

Dimensions of manipulability

Two important observations regarding the assessment of manipulability can be drawn from the previous chapter. First, our review suggests that its various components are best captured through a multidimensional approach – not only to distinguish structural and functional manipulability (SM and FM, respectively), but to also better characterise the latter. We have argued that a single scale is likely sufficient for SM. On the other hand, functional actions have properties relevant for different research questions that require separate assessments. The extant literature notably points to four candidate dimensions, namely to (i) a general evaluation of the association with manual actions for object use, (ii) the diversity of the typical uses, (iii) the distinctiveness of the actions, and (iv) the dissimilarity between use and transport actions. The second observation concerns methodological choices that can affect the reliability of the ratings. Our analyses indeed suggest that participants adjust their ratings relative to the items they have been presented. A failure to include a wide enough range of items from the two ends of the scale can thus result in inconsistent judgements. For manipulability ratings in particular, we have additionally seen that the instructions found in the literature often fail to mention the use of hands, the goal of the assessed actions, or both, thus likely introducing unnecessary ambiguity to their interpretation.

The aim of the current chapter is to present a new set of six manipulability ratings that incorporate the points highlighted above. To our knowledge, our dataset provides the most comprehensive assessment of manipulability to date, as well as the first manipulability norms for French words. Additionally, our discussions of the validity and usefulness of the instructions in the previous chapter were limited by the availability of – and the overlap between – reliable rating datasets. The present norms thus also extend our analyses by allowing a direct comparison of different manipulability ratings. We hope that our results will prove useful, as much for their practical use in experiments as for fostering further methodological inquiry.

5.1 The present ratings

As mentioned, a limitation of most manipulability instructions concerns their wording. To minimise ambiguities, those used in the current study explicitly stated that the actions being assessed are manual and specified their intended goal. The ratings were collected in three successive phases. First, SM ratings were obtained through the same task as Clarke and Lundington's (2018), i.e. by asking participants to rate the ease with which they could manually grasp an object in order to move it (*graspability*). The second task was a general assessment of FM that used a modified version of the 'action association' and 'hand/arm action association' instructions reviewed in Chapter 4 (Sections 4.3.2.1 and 4.3.2.2, respectively). Participants were asked to rate the extent to which they associate each word's referent with manual use (*action association*). A possible alternative

for this dimension was the 'hand necessity for function' instructions (Chapter 4, Section 4.3.2.3. Tranel et al., 1997) which also appear to yield relatively valid ratings. As was noted in our review, however, these require raters to consider each object's function, which can be ambiguous for some items (e.g. food, clothes).

The aim of the final rating study was to obtain a finer-grained assessment of FM on dimensions argued to be complementary in Chapter 4 (*functional attributes* hereafter). For each word, participants were first asked a binary question about whether it represents an object that can be used with the hands (*usability*). This step was introduced to draw their attention to manual functional actions specifically, and to avoid potentially disparate ratings for non-manipulable objects. As will be further explored in the results, the answers to this question were nevertheless also informative for the assessment FM. Following an affirmative response to an object's usability, participants were presented with three additional rating tasks. First, they estimated the number of manual actions they could typically perform with the object (number of actions – *NoA*). The NoA instructions were similar to those used by Lagacé et al. (2013), Guérard et al. (2015) and Heard et al. (2019), with the added precision that they concern manual actions. The second scale assessed the extent to which the hand posture for using each object differs from the one necessary to grasp and move it (*move-use dissimilarity*). The instructions were adapted from Lagacé et al. (2013) who similarly asked participants to rate "the extent to which the posture of the hand to use the object differed from the posture of the hand to grasp it" (p. 775–776). We nevertheless added the purpose of the grasp (i.e. to move the object) to stress that the two actions refer to different manipulations (see discussion in Chapter 4, Section 4.3.2.6). In the last rating task, participants were asked to assess the ease with which someone could recognise the object if they were to mime its typical use (*e*). As also discussed in Chapter 4 (Section 4.3.2.5), we used the pantomime task to capture the distinctiveness of FM as was originally intended by Magnié et al. (2003). Our instructions were similar to theirs, but slightly differed in their wording to highlight functional actions and to emphasize the ease of recognition instead of the ease to mime the object.

Dividing the rating tasks into three successive studies allowed us to present participants with a balanced set of items. In the graspability task, the stimuli were selected based on their body-object interaction ratings presented in Chapter 3. In turn, items in the action association questionnaire were drawn relative to their graspability scores, and those in the functional attributes questionnaire were contingent on their action association ratings.

5.2 Method

5.2.1 Participants

A total of 2454 participants responded to the rating questionnaires. 586 of them participated to the graspability rating task (431 female, 139 male, 10 'other' and 6 who did not wish to respond; Age: $Mdn = 22$, range = [18, 72]). A second group of 566 participants provided ratings for action association (398 female, 156 male, 9 'other' and 3 who did not wish to respond; Age: $Mdn = 27$, range = [18, 76]). Finally, 1302 participants (798 female, 482 male, 14 'other' and 8 who did not wish to respond; Age: $Mdn = 27$, range = [18, 81]) completed the questionnaire on functional attributes. All respondents were volunteers recruited mainly through social media platforms, except for 315 participants who participated through Prolific (all male; Age: $Mdn = 27$, range = [18, 50]) to the functional attributes questionnaire and received a compensation of £2.25. A description of the final sample of participants included in the computation of the ratings is provided in the data cleaning section below. All participants were over 18 years old and gave their consent at the beginning of the experiment.

5.2.2 Materials & Procedure

The stimuli used in all questionnaires were the same set of 1019 French words for which body-object interaction (BOI) ratings were previously collected (Chapter 3, Paisios et al., 2023). The three questionnaires were designed on Qualtrics and were administered separately, in the order presented below. The full instructions and scales can be found in Appendix C.

The target of the experiment was to obtain approximately 50 observations for each word and dimension. Words for which the target was reached were manually and incrementally removed from the sampling pool in the graspability task. An online document database (Cloud Firestore) was used for weighted stimulus sampling in the two other questionnaires. All questionnaires were designed so that they could be completed in approximately 10 to 15 minutes.

Graspability

In this task, participants rated on a 7-point scale the ease with which they can manually grasp what each word represents in order to move it. The scale ranged from 0 (impossible) to 6 (very easy), with individual labels describing each choice. A tick box labelled "N/A" was presented to the right of the scale to indicate when, and only when, a word was unknown. Participants were further asked, when possible, to interpret ambiguous words as physical objects.

Each participant was presented with 90 words[36] (experimental list hereafter) sampled from the full list of 1019 words. To ensure a balanced set of words regarding their graspability across participants, the sampling was based on the distribution of BOI ratings (Chapter 3, Paisios et al., 2023). BOI was divided into three categories, with ratings falling in [0 – 2) (low), (2 – 4) (midscale) and (4 – 6] (high). The experimental lists contained either 8 or 9 low-BOI words, 30 or 29 midscale-BOI words and 52 high-BOI words[37].

The experiment began with a consent form which participants had to confirm having read and accepted before continuing. They were then asked to complete a demographics questionnaire about their age, gender, education level and domain, handedness, whether they have a known language disorder, whether French is their native language or the age at which they have learnt it otherwise, the number of languages spoken in their parental home and whether French is part of them, whether they have lived in France during the first 18 years of their lives or the ages between which they have otherwise. Once completed, the full instructions of the rating task were presented on a new page and participants were asked to confirm having carefully read them. The same instructions were then repeated at the top of the next page, followed by the randomised list of 90 words and their respective rating scales to their right. The labels of the scale were repeated every 30 words.

Action Association

This task required participants to evaluate the extent to which each word's referent is associated to a use with the hands on a 7-point scale (0 – none, 6 – very strong). As for the graspability questionnaire, the scale options had individual labels and a "N/A" option was present to indicate not knowing a word. Besides the instruction to interpret, when possible, ambiguous words as physical objects, participants were additionally asked to only rate the presented words and not any potential associates (e.g. to not consider the actions performed with a *hammer* when rating the word *nail*).

The experimental lists consisted of 90 words, sampled based on their graspability ratings. As the variable displays a bimodal distribution (see section 5.4.4), a median split was performed to obtain two categories and 45 words were randomly drawn from each. The procedure was identical to the graspability rating task except for a slightly modified question in the demographic questionnaire. Instead of asking whether participants have lived in France for the first 18 years of their lives, the question more generally inquired about how many years they have lived in the country.

[36] Some participants were presented with slightly less words due to a technical error (87 words: $N = 2$; 88 words: $N = 4$; 89 words: $N = 25$).

[37] The number of sampled words was proportional to the total number of items in each category to ensure balanced sampling.

Functional Attributes

The functional attributes questionnaire required one or four evaluations on individually presented words. For each word, participants were first asked to indicate whether what it represents can be used with the hands (three options: 'Yes', 'No', 'I don't know this word'. *Usability* hereafter). If the given answer was 'Yes', participants were asked to further assess the word on three dimensions using 7-point scales. These were (i) the number of manual actions typically performed with the object ('0' to '6 or more'), (ii) the extent to which the posture of the hands to use the object differs from that necessary to grasp and move it (0 – identical, 6 – very different), and (iii) the ease with which someone looking at the participant miming the use of the object could recognise it (0 – impossible, 6 – very easy). If a word was judged as not referring to an object which can be used with the hands, it was automatically attributed a score of 0 on the three dimensions. As for the previous questionnaires, participants were asked to interpret the words as physical objects when possible. For the NoA ratings, they were further asked to base their judgments only on the actions most commonly associated with the objects, even if they could be used for other actions (e.g. screwing with a knife). The instructions for the pantomime ratings indicated that judgments should only be based on the ease to recognise the actual action performed during the use of the object, not a structural mime of it (e.g. miming the use of scissors using two fingers to represent the object).

All participants were first presented with 6 words that were not in the general stimulus pool in random order to serve as anchors (*blue*, *bracelet*, *corkscrew*, *dust*, *hole punch*, *partition*), followed by 45 words sampled from the same list used in the previous questionnaires. The sampling was based on the action association ratings divided into three quantiles, with 15 words randomly drawn from each.

As with the two previous questionnaires, participants started by giving their consent and by providing demographic information (identical to the action association questionnaire). They were then presented with a short summary of the procedure, followed by the full instructions to the three functional attributes questions (i.e. NoA, pantomime, move-use dissimilarity). Participants were further informed that they can have access to the instructions at any time through pop-ups that appeared by clicking on the title of each question.

5.2.3 Data analysis and availability

All data wrangling and analyses were performed with R (v4.3.2, R Core Team, 2023) and through the RStudio interface (v2023.12.1.402, RStudio Team, 2024). The trial-level data (raw and

pre-processed), summary statistics and scripts are available upon request and will be shared in open access when the current work is ready for publication.

5.3 Results

5.3.1 Data cleaning

Multiple criteria were used to clean the data. These somewhat differed for the functional attributes questionnaire due to its different procedure and are presented separately. For the graspability and action association questionnaires, we started by removing participants who rated less than two thirds of the items (Graspability: $N = 9$; Action association: $N = 9$). Those who reported having learnt French after the age of two and to have lived in France for less than 10 years (Graspability: $N = 9$; Action association: $N = 12$), or to have learnt French before the age of two but to have lived in France for less than five years (Graspability: $N = 8$; Action association: $N = 7$) were further removed from the final dataset. Next, participants who had an inter-item standard deviation (ISD. See Chapter 3 and Marjanovic et al., 2015) lower than 2.5 standard deviations from the group's average were dropped from the analyses (Graspability: $N = 10$; Action association: $N = 8$). The final step screened participants for their deviation from the group (*deviation analysis*). For each respondent, we computed the corrected[38] average ratings and agreement rates (i.e. maximum proportion of responses within a 3-unit interval) for the items they had rated. The agreement rates were then rescaled from [.4, 1] to [0, 1] to serve as weights in three analyses. We inspected the weighted Pearson correlation (r_w), the weighted linear regression coefficient (B_w) and the weighted median difference (Mdn_w) between each participant's ratings and the group's corrected average[39]. Items detected as having a bimodal response distribution (see procedure in Chapter 3) were not included in these analyses. Participants with $r_w < 0.5$, $B_w \notin [0.5, 1.5]$ or a Mdn_w above 2.5 standard deviations from the group's average were removed from the dataset (Graspability: $N = 9$; Action association: $N = 32$).

In the functional attributes questionnaire, participants who rated less than two thirds of the items ($N = 2$) and who did not meet the same language criteria described above ($N = 15$) were similarly removed in the first step. One additional participant was dropped for not indicating the number of years they have lived in France. Next, $N = 28$ participants who did not judge a *corkscrew* (anchor word common to all participants) as usable with the hands were removed. Two

[38] By 'corrected' we mean that the summary statistics were computed after removing a given participant's responses.
[39] For the linear regression, participants' ratings were entered as the predictor variable.

additional screenings were performed on usability judgements. The first was the same as the deviation analysis described above, with the dichotomous Yes/No responses treated as numerical (1 and 0 respectively). The weights were the corrected maximum proportion of 'Yes' or 'No' responses rescaled from [.5, 1] to [0, 1]. In the second screening, we filtered the items in each participant's list judged as either usable or non-usable by at least 90% of the rest of the group. Respondents who did not provide matching judgements for at least two thirds of either category were flagged as outliers. These two methods led to the removal of $N = 92$ unique participants. The following step involved an analysis of ISDs similar to the one described above. Each participant's ISD was determined based on the items judged as usable and by combining the ratings to the three functional scales. Respondents with an ISD below 2.5 standard deviations from the group's average were removed from the dataset ($N = 9$). Finally, we examined the deviation of participants' responses from the corrected group averages for the NoA and pantomime ratings, after filtering the items with an agreement above 80% regarding their usablility. For NoA, only the weighted median was used as a criterion due to a low number of items on the high end of the scale, with the threshold being 2.5 standard deviations above the group's average median difference. The weights were derived from the usability judgements as described above. The screening of pantomime ratings was identical to the deviation analysis presented for the two previous questionnaires. These analyses led to the removal of $N = 40$ unique participants. The screening procedure did not include the move-use dissimilarity ratings as these display a large number of disagreements and included very few items on the high end of the scale (see descriptive analyses, Section 5.4.4).

The final datasets were composed of 541 eligible participants for the graspability questionnaire (398 female, 129, male, 9 'other', 5 who did not wish to respond. Age: $Mdn = 22$, range = [18, 72]), 494 respondents for the action association questionnaire (343 female, 139, male, 9 'other', 3 who did not wish to respond; Age: $Mdn = 28$, range = [18, 76]) and 1115 participants for the functional attributes questionnaire (664 female, 431 male, 13 'other', 7 who did not wish to respond; Age: $Mdn = 27$, range = [18, 79]). The total number of valid ratings and descriptive statistics for the word-wise number of observations are presented in Table 8. All ratings were derived by averaging the responses for each item. For usability, the dichotomous Yes/No judgements were treated as numerical (respectively 1 and 0). Their average is thus equivalent to the proportion of 'Yes' responses.

5.3.2 Reliability metrics

The internal reliability of the present norms was assessed through the same metrics used for the body-object interaction ratings presented in Chapter 3. The summary statistics for the three analyses are given in Table 9. Person-total correlations indicate the extent to which, on

Table 8
Summary of the total and item-wise number of observations for each questionnaire

	Total number of ratings	'I don't know this word' responses	Item-wise observations		
			M	*SD*	range
Graspability	48150	507	47.2	1.8	[41, 54]
Action association	44321	139	43.5	2.5	[36, 53]
Functional attributes	225552	477	48.8	2.7	[40, 58]

average, the ratings of participants correlated with the corrected group averages. The mean average absolute z scores represent the absolute difference between participants' responses and the group's average ratings in standard units. Finally, the split-half reliabilities were obtained by averaging the Spearman-Brown corrected correlations between the mean ratings of randomly split halves of the sample over 1000 iterations. The metrics generally point to a good reliability of the norms and were comparable to those reported by other rating studies (e.g. Desrochers & Thompson, 2009; Paisios et al., 2023; Pexman et al., 2019; Stoinski et al., 2023). We can nevertheless note the relatively low person-total correlations for the NoA and move-use dissimilarity ratings. In the former case, this might be partly due to the ambiguity of our instructions regarding what constitutes a valid action when rating a given item. Most NoA ratings were additionally clustered on the low end of the scale (see descriptive analyses below). Variations in individual ratings could thus have affected the strength of the correlations. For the move-use dissimilarity ratings, the low correlations likely stemmed from very few items being rated on the high end of the scale and from a large portion of them generating disagreements.

Table 9
Results of the internal reliability analyses

	Person-total correlation	Mean average absolute z score	Split-half reliability
Ease to grasp	0.84 (0.06)	0.69 (0.15)	0.99 (0.00)
Action association	0.74 (0.09)	0.77 (0.15)	0.98 (0.00)
Usability	0.77 (0.09)	0.67 (0.16)	0.98 (0.00)
Number of actions	0.65 (0.12)	0.67 (0.23)	0.96 (0.00)
Ease to pantomime	0.78 (0.08)	0.70 (0.17)	0.98 (0.00)
Move-Use dissimilarity	0.61 (0.13)	0.74 (0.16)	0.96 (0.00)

Note. Standard deviations are reported in parentheses.

5.3.3 Comparison with published datasets

The present data were also compared to all the available manipulability ratings identified in Chapter 4. The relationship between each pair of ratings was investigated through a Pearson correlation and a simple linear regression (Table 10). In the latter case, the present norms were entered as the predictor variable. The rationale for inspecting the regression coefficients was that correlations are only informative about the strength and direction of the relationship, but not the rate of change (i.e. the slope). If two sets of ratings capture the same dimension, then both coefficients should approach 1. To illustrate the difference, the correlation between our and Heard et al.'s (2019) graspability ratings was $r(160) = .90$, $p < .0001$. However, the slope of the relationship was only $B = 0.59$, $p < .0001$. The reason for this divergence is that most items tended towards the middle of the scale in Heard et al.'s (2019) datasets, whereas they were distributed closer to the ends of the scale in the present work.

The results overall suggest that our graspability, action association, usability and pantomime ratings were largely consistent with the literature and with the concerns raised about some of the datasets in Chapter 4 (see Appendix B in particular). The graspability ratings were most strongly related to the available SM variables, and especially to the ease to grasp (Guérard et al., 2015), the likelihood to grasp (Amsel et al., 2012; Díez-Álamo et al., 2018) and the ease to hold (Stoinski et al., 2023). The action association dimension displayed stronger relationships with FM variables such as hand/arm action association (Binder et al., 2016) and hand necessity for function (Moreno-Martínez & Montoro, 2012; Moreno-Martínez et al., 2011, 2014; Miklashevsky, 2018; Navarrete et al., 2019) variables. They were also highly correlated with both versions of the pantomime ratings[40] (Guérard et al., 2015; Magnié et al., 2003), suggesting that the results obtained through the modified instructions (Guérard et al., 2015) did not strongly differ from those using the original instructions (Magnié et al., 2003). Usability judgements displayed a highly similar pattern of results, but with somewhat weaker correlations and especially lower regression coefficients. The reason for this was that most Likert-type ratings contained a significant number of midscale items, while the binary usability judgements yielded a highly bimodal distribution (Figure 15). As will be further discussed below, the latter dimension thus appears to have better discriminated FM from non-FM objects. Our pantomime ratings were most strongly related to those provided by Magnié et al. (2003) and Guérard et al. (2015), and to a slightly lower extent to the same variables as the action association ratings.

[40]The original pantomime instructions refer to those used by Magnié et al. (2003) which served as the basis for the current study. The modified instructions introduced by Guérard et al. (2015) did not assess how recognisable the mimed actions would be, instead focusing simply on the ease to mime them.

Our ratings for move-use dissimilarity and for NoA in particular diverged from the available norms for these variables. In the former case, the correlations were moderate with both Salmon et al.'s (2010) and Lagacé et al.'s (2013) ratings. We can nevertheless note the relatively low number of items in common and that the slope of the relationship with Lagacé et al.'s (2013) ratings was much closer to 1 compared with Salmon et al.'s (2010) dataset. The instructions used for this dimension in the current study were indeed much more similar to those used by Lagacé et al. (2013). Additionally, a direct comparison of these two datasets showed that their ratings similarly displayed a moderate correlation ($r(56) = .46$, $p = .0003$). These results, along with the descriptive analyses below, suggest that the move-use dissimilarity instructions generate significant disagreements in the ratings and that they might thus not be very reliable. For NoA, both the correlation and the regression coefficients with the three available datasets were generally low (Guérard et al., 2015; Heard et al., 2019; Lagacé et al., 2013). A potential reason for this result is that our instructions were primed by the usability judgements and specifically assessed the number of *manual* actions, whereas none of the three studies included this precision. As we have previously mentioned, it is also possible that it stems from the NoA ratings being generally low or from the inherent ambiguity of the instructions as they do not clearly define how actions should be distinguished and counted. Unfortunately, the studies reporting NoA ratings have too few items in common to allow a meaningful analysis of their correlations with one another for comparison.

Table 10

Pearson correlation and linear regression coefficients between the current ratings and all other available datasets providing manipulability ratings

			Graspability		Action association		Usability		Number of actions		Pantomime		Move-Use dissimilarity	
Study	Variable	*N*	*B*	*r*	*B*	*r*	*B*	*r*	*B*	*r*	*B*	*r*	*B*	*r*
a	Ease to grasp	304	0.94	0.83	0.72	0.56	0.69	0.65	1.54	0.56	0.71	0.56	0.75	0.43
b	Ease to grasp	162	0.59	0.90	0.54	0.64	0.47	0.73	1.13	0.66	0.52	0.62	0.55	0.53
c	Ease to grasp	514	0.70	0.82	0.55	0.60	0.49	0.66	1.13	0.56	0.45	0.52	0.51	0.44
d	Likelihood to grasp	259	0.92	0.90	0.74	0.63	0.67	0.69	1.60	0.63	0.64	0.55	0.72	0.47
e	Likelihood to grasp	353	0.92	0.92	0.74	0.61	0.71	0.74	1.67	0.66	0.68	0.57	0.91	0.58
c	Ease to hold	514	0.90	0.87	0.62	0.55	0.57	0.62	1.30	0.52	0.51	0.47	0.57	0.39
a	Ease to move	304	0.86	0.84	0.62	0.54	0.60	0.63	1.34	0.54	0.61	0.53	0.68	0.43
c	Ease to move	514	0.74	0.81	0.51	0.53	0.47	0.59	1.11	0.51	0.43	0.46	0.47	0.37
f	Grasp for moving	212	0.40	0.42	0.22	0.32	0.25	0.30	0.23	0.16†	0.16	0.25	ns	ns
g	Action association	109	0.16	0.30	0.42	0.70	0.31	0.61	0.79	0.56	0.43	0.71	0.47	0.56
h	Hand/arm association	142	0.38	0.49	0.83	0.82	0.48	0.63	1.31	0.61	0.66	0.68	0.75	0.61
i	Hand/arm association	929	0.21	0.38	0.52	0.69	0.27	0.51	0.75	0.50	0.41	0.55	0.38	0.44
j	Hand/arm association	230	0.33	0.50	0.66	0.74	0.37	0.59	1.05	0.60	0.51	0.59	0.50	0.48
k	Manipulation	447	0.22	0.49	0.41	0.72	0.26	0.62	0.66	0.58	0.37	0.66	0.42	0.61
l	Hand necessity	51	ns	ns	0.99	0.83	0.72	0.72	2.32	0.73	0.93	0.76	1.12	0.74
m	Hand necessity	188	0.47	0.51	0.85	0.81	0.64	0.76	1.56	0.69	0.71	0.71	0.97	0.72
n	Hand necessity	246	0.42	0.52	0.79	0.81	0.53	0.72	1.58	0.71	0.67	0.70	0.78	0.63
o	Hand necessity	174	0.36	0.52	0.70	0.80	0.52	0.76	1.39	0.71	0.62	0.72	0.83	0.68
p	Hand necessity	163	0.48	0.51	0.94	0.88	0.68	0.79	1.86	0.74	0.79	0.76	1.00	0.73
q	Pantomime (original)	216	0.40	0.47	0.74	0.79	0.54	0.72	1.34	0.63	0.77	0.85	0.76	0.63
r	Pantomime (original)	212	ns	ns	0.35	0.40	0.27	0.24	ns	ns	0.52	0.63	0.19	0.15†
s	Pantomime (original)	200	0.29	0.21	0.53	0.59	0.54	0.43	ns	ns	0.66	0.79	0.31	0.22‡
t	Pantomime (original)	313	0.13	0.21	0.45	0.63	0.26	0.47	0.60	0.37	0.41	0.61	0.43	0.51
a	Pantomime (modified)	304	0.49	0.49	0.87	0.77	0.56	0.60	1.23	0.51	0.81	0.73	0.73	0.47
f	Pantomime (modified)	212	0.41	0.38	0.28	0.36	0.36	0.37	ns	ns	0.32	0.44	0.30	0.25
b	Pantomime (modified)	162	ns	ns	0.24	0.40	0.12	0.27	0.31	0.25‡	0.31	0.52	0.19	0.25‡
u	Grasp/Use dissimilarity	117	ns	ns	0.20	0.29‡	0.16	0.22†	ns	ns	0.22	0.35	0.50	0.49
v	Grasp/Use dissimilarity	83	ns	ns	ns	ns	ns	ns	0.49	0.25†	ns	ns	0.91	0.57
v	Number of actions	92	ns	ns	0.22	0.34‡	0.18	0.22†	0.39	0.31‡	0.21	0.33‡	ns	ns
a	Number of actions	304	0.06	0.19	0.13	0.36	0.06	0.18‡	0.22	0.27	0.09	0.26	0.12	0.23
b	Number of actions	162	ns	ns	ns	ns	ns	ns	0.17	0.22‡	ns	ns	ns	ns
u	Grasp and/or use	200	1.07	0.93	0.87	0.66	0.79	0.73	1.92	0.65	0.80	0.62	0.89	0.52
w	Grasp and/or use	249	0.74	0.72	0.86	0.87	0.80	0.92	2.14	0.82	0.81	0.81	0.99	0.73

Notes. The current ratings were treated as predictor variables in the linear regression analyses.
ns: non-signficant; † $0.05 > p \geq 0.01$; ‡ $0.01 > p \geq 0.001$; all other $ps < .0001$.
The colour code represents ranges of 0.2 for the Pearson correlation coefficients.
a: Guérard et al. (2015); b: Heard et al. (2019); c: Stoinski et al. (2023); d: Amsel et al. (2012); e: Díez-Álamo et al. (2018); f: Clarke & Lundington (2018); g: Hoffman & Lambon Ralph (2013); h: Binder et al. (2016); i: Lynott et al. (2020); j: Repetto et al. (2023); k: Medler et al. (2005); l: Moreno-Martínez et al. (2011); m: Moreno-Martínez & Montoro (2012); n: Moreno-Martínez et al. (2014); o: Miklashevsky (2018); p: Navarrete et al. (2019); q: Magnié et al. (2003); r: Brodeur et al. (2010); s: Brodeur et al. (2012); t: Brodeur et al. (2014); u: Salmon et al. (2010); v: Lagacé et al. (2013); w: Ni et al. (2019).

5.3.4 Descriptive analyses

The descriptive statistics for the present ratings and for the interrater agreement scores for each dimension (except for usability and NoA) are provided in Table 11. As with the body-object interaction ratings presented in Chapter 3, the agreement scores reflect the highest proportion of responses within any consecutive 3-unit interval. They were not computed for usability and NoA as these two dimensions were not rated on unipolar Likert-type scales (respectively binary and numerical). In what follows, we will refer the high and low end of the scale will respectively be used to refer to the higher and lower third portions of the scale. Keeping in line with our analyses in the previous chapter, we will consider that items with an interrater agreement rate below .65 display significant disagreement.

Table 11
Descriptive statistics for the present ratings and for their agreement rates (N = 1019)

	M	*SD*	*Min*	1st *Q*	*Mdn*	3rd *Q*	*Max*	Skewness	Kurtosis
Graspability [0, 6]									
Ratings	3.45	2.13	0.00	1.35	3.93	5.57	6.00	– 0.30	– 1.47
Agreement	0.85	0.15	0.43	0.74	0.91	0.98	1.00	– 0.88	– 0.46
Action association [0, 6]									
Ratings	2.17	1.58	0.04	0.77	1.89	3.29	6.00	0.53	– 0.82
Agreement	0.77	0.16	0.43	0.62	0.78	0.93	1.00	– 0.14	– 1.30
Usability [0, 1]									
Ratings	0.49	0.38	0.00	0.09	0.48	0.87	1.00	0.01	– 1.61
Number of actions [0, 6]									
Ratings	0.95	0.79	0.00	0.17	0.90	1.55	4.33	0.53	– 0.25
Ease to pantomime [0, 6]									
Ratings	1.60	1.59	0.00	0.21	1.06	2.64	5.92	0.88	– 0.36
Agreement	0.83	0.15	0.46	0.70	0.87	0.98	1.00	– 0.56	– 0.95
Move-Use dissimilarity [0, 6]									
Ratings	1.71	1.36	0.00	0.35	1.70	2.81	5.52	0.31	– 1.07
Agreement	0.75	0.18	0.42	0.58	0.73	0.94	1.00	0.03	– 1.46

The frequency distributions and the relationship of the SDs to the average ratings – except for usability – are presented in Figure 15. The plots show clear bimodal distributions for graspability and usability, indicating that participants generally agreed in their judgements. A few items towards the middle of the graspability scale additionally had relatively high agreement rates. A closer inspection revealed that these were mostly animals (e.g. *donkey*, *monkey*, *turkey*), furniture (e.g. *chest*, *desk*, *sofa*) and appliances (e.g. *refrigerator*, *television*, *radiator*). However, several animals were also disagreed on (e.g. *flea*, *fish*, *owl*, *squirrel*), along with body parts (e.g. *elbow*, *kidney*,

skin) and human-related words (e.g. *clown*, *uncle*, *woman*). The graspability instructions used in the current work thus appear to have resulted in highly comparable ratings to the other SM dimensions found in the literature

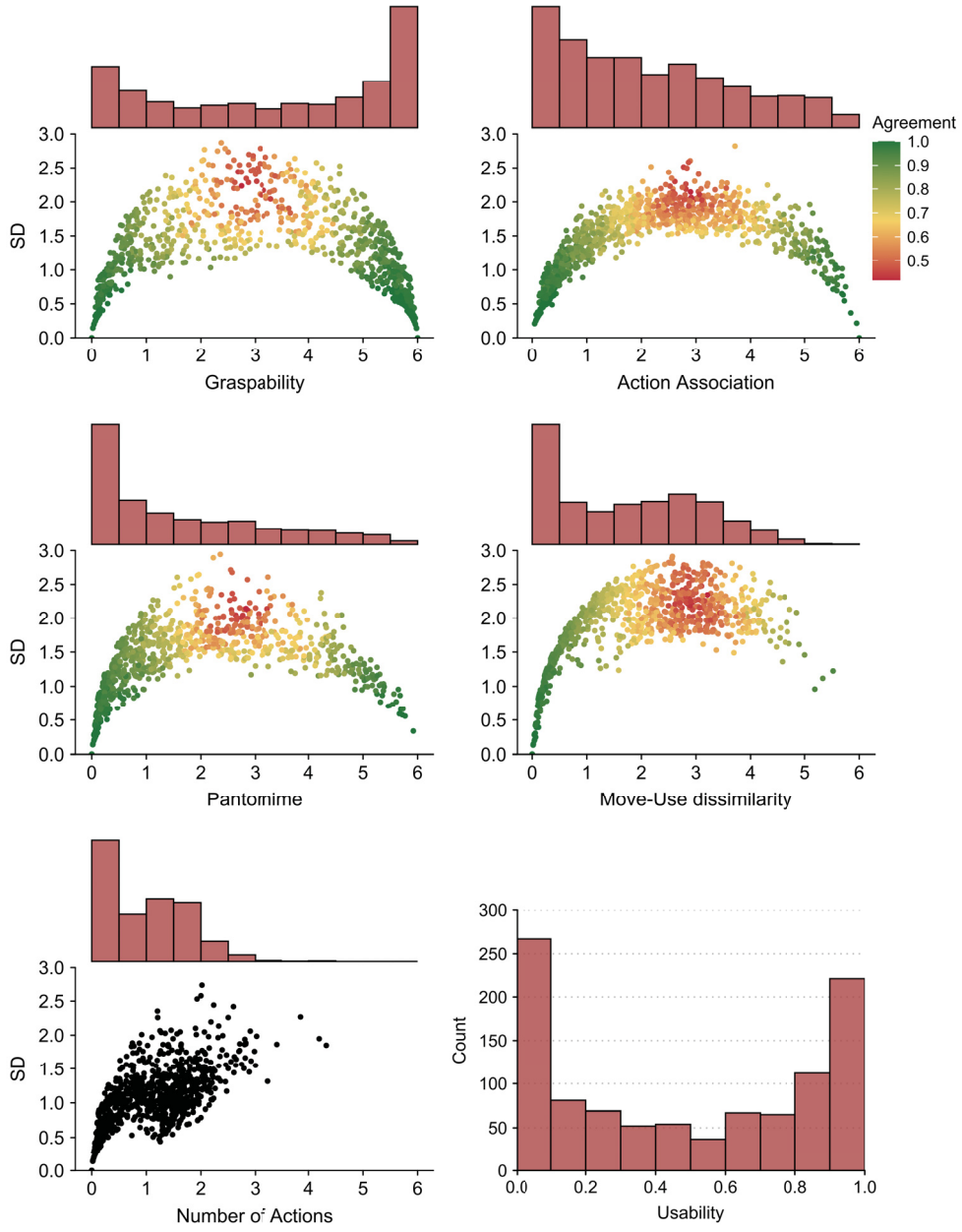
The action association and move-use dissimilarity ratings displayed strongly skewed distributions. These norms notably had a large number of items for which participants disagreed (respectively 30% and 37%). In the former case, these were mostly clothes (e.g. *hat*, *scarf*, *vest*), foods (e.g. *almond*, *garlic*, *soup*), furniture (e.g. *cupboard*, *dresser*, *shelf*), appliances (e.g. *oven*, *stove*, *freezer*) and vehicles (e.g. *car*, *motorcycle*, *sailboat*), along with some manmade manipulable objects (e.g. *diary*, *envelope*, *microphone*). This shows that explicitly stating the action's goal in the instructions did not reduce their ambiguity as much as we had hoped in light of our discussions in the previous chapter. It should be noted that the high end of the scale nevertheless contained almost exclusively FM objects (e.g. *fork*, *comb*, *hammer*, *pencil*, *piano*, *spear*), with the only exceptions being a few words referring to the hand (e.g. *finger*, *slap*, *thumb*). The modified action association instructions thus appear to reliably detect most FM items on the high end of the scale despite several object categories being disagreed on. In contrast, only very few items were reliably rated high for their move-use dissimilarity, none of which had an agreement rate of 1 (e.g. *chair*, *door*, *piano*, *printer*, *oven*). Those that were disagreed on were relatively much more heterogeneous, and critically included a large number of FM objects (e.g. *calculator*, *cigarette*, *fork*, *lighter*, *needle*, *stapler*). Our modification of the instructions thus does not appear to have reduced their ambiguity in any meaningful way, and our move-use dissimilarity ratings cannot be reliably considered to capture objects that differ in the way they are structurally and functionally manipulated.

The pantomime ratings similarly displayed a highly skewed distribution, with nevertheless relatively fewer items generating disagreement (17%). These were generally quite heterogeneous, notably including several large household items (e.g. *dresser*, *oven*, *table*, *wardrobe*, *television*), some foods (e.g. *apple*, *bread*, *pancake*, *pill*) and wearable items (e.g. *brooch*, *coat*, *shirt*, *shoe*). The high end of the scale was mostly composed of FM objects such as tools (e.g. *axe*, *hammer*, *screwdriver*), instruments (e.g. *guitar*, *piano*, *violin*), weapons (e.g. *bow*, *gun*, *sword*), and other FM items (e.g. *comb*, *knife*, *paintbrush*, *racket*). Overall, our ratings appear to be highly similar to those provided by Magnié et al. (2003). Having specifically targeted the pantomime of functional manual actions nevertheless appears to have removed some ambiguity, notably for body parts that were rated low in the current norms (e.g. *arm*, *hand*, *nose*).

Usability and NoA ratings are a little less straightforward to interpret as an interrater agreement rate could not be computed. In the former case, relatively few (16%) and highly heterogeneous items can be found in the middle third of the scale. Among these we can notably note a large number of edible items (e.g. *garlic*, *honey*, *quiche*, *salad*) and animals (e.g. *chicken*, *crab*, *fish*,*lobster*). One possibility is that these animal names are ambiguous as they can also interpreted as food. However, the presence of typically non-edible animals (e.g. *cat*, *dog*, *puppy*) does not directly allow this

Figure 15

Standard deviations as a function of the average ratings and frequency distributions for the present norms



Note. The colour code for the first four plots represents the item-level interrater agreement scores.

interpretation and it is not entirely clear why participants disagreed about whether they can be used with the hands. As previously noted, the distribution of NoA ratings showed that the vast majority

of the items were found on the low end of the scale, with only a handful having relatively higher ratings (e.g. *balloon*, *telephone*, *thread*, *toy*, *rope*). We can additionally note the vast disparity in SDs for the items with an average rating between 1 and 2. These observations strongly suggest that the NoA instructions do not allow to reliably capture the diversity of the actions associated with objects, most likely because they are ambiguous as to what can be counted as an action.

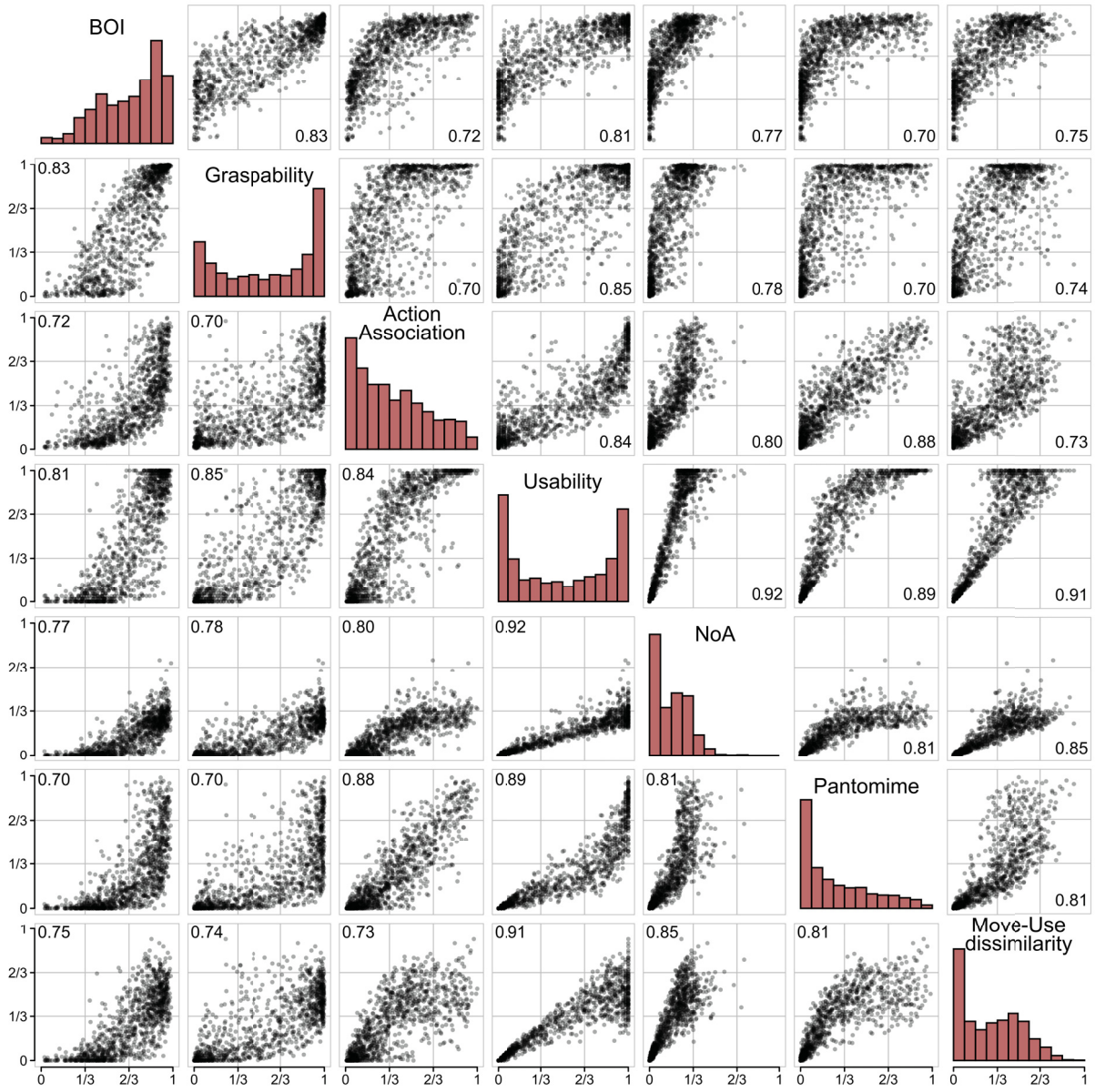
5.3.5 Comparative analyses

Pair-wise comparisons of all the ratings collected in the current work are shown in Figure 16. The body-object interaction (BOI) ratings presented in Chapter 3 were also included to allow a direct assessment of their relationship with the other manipulability variables. We can see that items rated high for their graspability, usability, action association and pantomime also received almost exclusively high ratings for their BOI. Items rated low on the FM-related dimensions (usability, action association, pantomime) nevertheless also had highly variable BOI ratings across its entire range. In contrast, items rated low for their graspability were generally restricted to low and midscale BOI values. We can additionally observe that a large number of words were found in the middle of the scales for both of these variables, suggesting that there was some overlap in the type of items that generated disagreements with both instructions. Such a pattern was much less apparent in the comparisons of BOI to the FM-related variables, which displayed much stronger non-linear trends. In line with Heard et al.'s (2019) analysis (see Section 2.4), these observations suggest that BOI is indeed very closely related to manual interaction and to graspability in particular. Note, however, that the different patterns of disagreement (midscale) in these variables also suggest that BOI cannot be entirely reduced to manipulability.

Starting with the FM scales suspected to be largely uninformative following the descriptive analyses (i.e. NoA and move-use dissimilarity), their comparison with the rest of the variables suggests that they were indeed not sensitive enough to detect the intended dimensions. The main purpose of the NoA ratings was to assess the diversity in the actions associated with FM objects. As can be seen on the plots, however, NoA values were highly similar for items found on both the middle and the upper end of the action association scale. The comparison with usability reveals a much stronger linear relationship and slightly more variability for items consistently judged as usable. However, the extreme clustering of NoA ratings does not allow to reliably differentiate objects associated with more actions from others. The failure to observe a similar pattern with the action association ratings additionally suggests that the few items rated slightly higher for their NoA might not be particularly associated with FM (see also the examples provided in the descriptive analyses). Regarding the move-use dissimilarity ratings, the comparisons reveal that the majority of items rated high or midscale for their usability or for their action association generated disagreements on this dimension. Coupled with the elements provided in the descriptive analyses,

Figure 16

Correlation plot for the current and for our body-object interaction ratings



Note. BOI: body-object interaction; NoA: number of actions.
All variables are rescaled to [0, 1] for better readability.
Pair-wise Pearson correlation factors are provided in the corners of the scatter plots. All $ps < .0001$.

this largely confirms that the move-use dissimilarity ratings mainly resulted in disagreements for a large number of items, except for those consistently detected as not associated with FM.

Considering midscale items on the usability scale to be a result of disparity in the judgements, Figure 16 shows that the usability and action association instructions were highly similar, but that they also yielded different patterns of disagreements. Interestingly, most midscale usability items were rated low for their action association, whereas midscale action association ones were generally judged as usable. In the former case, these mostly referred to animals (e.g. *carp*, *chicken*,*snake*), plants (e.g. *grass*, *jasmine*, *wheat*), furniture and some other household objects (e.g. *armchair*, *carpet*, *mattress*, *spotlight*). It thus appears that such items generated ambiguity regarding their possibility to be manually used, whereas their association to manual uses was rated as low regardless. In contrast, a large number of items were disagreed on regarding their association to manual uses but were mainly judged as usable. Among these were notably foods (e.g. *biscuit*, *cake*, *carrot*, *cheese*, *grape*), furniture and appliances (e.g. *chair*, *chest*, *fridge*, *stove*, *wardrobe*), some manipulable items (e.g. *bell*, *box*, *clothes hanger*, *jar*, *kettle*) and wearable items (e.g. *boot*, *brooch*, *cap*, *helmet*, *ring*,*tie*). We can additionally note that a few items received high ratings for usability but were rated low for action association (e.g. *diamond*, *filter*, *poster*, *stool*), while other received high (*palm*, *slap thumb*) or midscale (e.g. *beard*, *dance*, *garden*, *kitchen*) ratings on the latter scale but were judged as not usable. Overall, these results suggest that usability judgements reflect a rather generic and too general assessment of FM, and that the action association ratings are more closely related to actual use – as would be expected. The action association dimension nevertheless also led to relatively more disagreements, likely due to ambiguities that emerged in relation to some words. Given the highly overlapping, but slightly different and complementary results obtained through these two instructions, using both would arguably result in the most reliable set of FM and non-FM items.

The comparison of the two dimensions discussed above with the graspability ratings is consistent with the relationship that would have been expected between FM and SM variables. Most items rated low on graspability also had low action association values, except for few that were on the high end (e.g. *dispenser*, *piano*, *slap*) or the middle of the scale (e.g. *car*, *hunting*, *kitchen*, *sport*, *stove*). Very similar results can be observed with the usability judgements. For graspable objects, action association ratings and, to a somewhat lesser extent, usability judgements were much more scattered across their respective scales. This shows that the FM and SM variables indeed captured different dimensions, as graspable objects generally tend to be associated with manual actions (e.g. *camera*, *hammer*, *knife*, *joystick*, *scissors*), but are not necessarily always so (e.g. *fish*, *grass*, *picture*, *seaweed*, *shell*, *tooth*).

Finally, the pantomime ratings were surprisingly similar to the action association ones, with nevertheless some small differences. First, we can see that some items found in the middle of the action association scale were more consistently rated low for pantomime. Most of these were foods (e.g. *artichoke*, *butter*, *fries*, *mango*, *onion*, *pie*), along with a few other words from different categories (e.g. *cat*, *dance*, *foam*, *kitchen*, *tractor*, *tube*, *wood*). Conversely, some items rated high for their action association generated disagreements with the pantomime instructions (e.g.

blender, *drawer*, *highlighter*, *pliers*, *spatula*, *tape*, *toy*). Pantomime ratings thus appear to 'resolve' some of the ambiguity in the action association instructions by leading to lower ratings, but to also miss some FM objects. Interestingly, there were a few items rated high for pantomime but found midscale in action association. These were mostly wearable (e.g. *belt*, *crown*, *earphones*, *scarf*, *tie*, *wristwatch*) and a few other objects (e.g. *bell*,*car*, *chair*, *microphone*, *motorcycle*) that can be considered to be FM in some cases but appear to have generated disagreements with the latter instructions. For pantomime ratings to be a useful dimension, they would have arguably needed to dissociate objects associated with a recognisable action from those associated with a less distinguishable one. We can nevertheless see that virtually no items were rated high for their action association and low for pantomime, with the few exceptions being mostly hand-related words (e.g. *finger*, *fist*, *hand*, *palm*). Thus, the pantomime ratings do not appear to be a particularly valid assessment of the distinctiveness of the actions associated with objects, but rather a variation of the most generic action association ratings.

5.4 Discussion

The goal of the current chapter was to present a new set of manipulability norms for words to allow their availability in French and to extent our investigations into their validity. The data were collected in three studies and through six different rating instructions inspired by our review in Chapter 4 that were argued to be the most appropriate to assess structural manipulability and various aspects of functional manipulations (SM and FM, respectively). We were particularly attentive to the wording of the instructions so that they explicitly stated the goal of the actions being rated, and that only manual actions were concerned. As discussed in Chapter 4, we hoped that this would solve some of the disagreements identified in the published datasets that used similar instructions. The first study assessed SM with a single scale on the ease with which an object could be grasped in order to be moved (*graspability*, Clarke & Lundington, 2018). A general assessment of FM was then performed by asking participants to rate the extent to which they associate each object with manual use (*action association*). The goal of the third and final study was to capture more specific characteristics of FM that have been previously used in the literature. Participants were first presented with a binary yes/no question about whether a given word represents an object that can be manually used (*usability*). Following an affirmative response, they were presented with three rating scales respectively assessing (i) the number of manual actions they could typically perform with the object (*number of actions* – NoA, Lagacé et al., 2013), (ii) the extent to which their hand posture for using it would differ from the one necessary to grasp and move it (*move-use dissimilarity*, Lagacé et al., 2013), and (iii) the ease with which someone could recognise the object if they were to mime its typical use (*pantomime*, Magnié et al., 2003).

The ratings were submitted to several analyses to assess their reliability and their validity. First, standard reliability analyses alerted us on potential issues with two of the norms in particular, namely NoA and move-use dissimilarity. Note that these metrics were generally difficult to interpret, a point to which we will return in the General Discussion. A correlational approach was then used to compare our ratings to those provided in published datasets. The results showed that our graspability, usability, action association and pantomime norms were generally consistent with related variables found in the literature. However, the NoA and move-use dissimilarity ratings had much lower correlations with the available datasets than what would have been expected. Further descriptive analyses informed by the midscale disagreement problem (see Chapters 3 and 4) revealed that the move-use dissimilarity instructions generated significant disagreements and that they did not reliably capture any meaningful items on the high end of the scale. For NoA, most items were clustered on the low end of the scale, rendering the ratings largely uninformative due to their low discriminatory power. In contrast, the four other scales were generally found to capture their intended dimensions. The disagreement patterns observed across different item categories were nevertheless quite similar to those of other datasets discussed in Chapter 4. This suggests that explicitly stating manual actions and their goals in the instructions did not reduce their ambiguity as much as we had hoped. It should nevertheless be noted that these observations relied on manual inspections of the ratings and that more systematic analyses as those presented in Chapter 4 could reveal differences that we have missed. Unfortunately, we did not have access to such data and our attempt to categorise our words through large language models proved largely unsatisfactory due to a large number of errors.

To better assess the validity of the ratings and what they capture, we finally performed pairwise comparisons between all of the variables presented in the current chapter, as well as with the body-object interaction (BOI) ratings introduced in Chapter 3. For the latter variable, we found that BOI closely – but not entirely – reflected manual interactions with objects, thus corroborating Heard et al.'s (2019) findings. Overall, the graspability, usability and action association ratings displayed coherent relationships. Our results suggest that the usability judgements captured the possibility, in principle, of using an object, thus serving as a rather generic assessment of FM. However, this also resulted in several items not particularly associated with FM to receive large values. In contrast, the action association ratings displayed a much more graded distribution and appear to have been more sensitive to actual use. This scale nevertheless also yielded significant disagreements for a large number of items, as well as a few items whose ratings are rather inconsistent with the assessment of FM. As these two dimensions generated disagreements for different and complementary types of items, the most reliable categorisation of objects as FM and non-FM would likely be obtained by considering both. For the remaining variables, we surprisingly found that the pantomime ratings were very closely related to the action association ones and that they were not particularly sensitive to the distinctiveness of FM – except for items that generated disagreements on the latter scale. This dimension should thus be considered a general assessment of FM and does not appear to bring

much added value as a complementary variable. It should nevertheless be noted that a large number of foods (particularly fruits and vegetables) were rated low for their pantomime but generated disagreements with the action association instructions. Using pantomime ratings would thus result in categorising them as non-FM objects, whereas they can arguably be expected to be associated with FM to some extent. Finally, the relationships of the NoA and move-use dissimilarity ratings to the other variables largely confirmed that they were not sensitive enough to reliably capture a meaningful aspect of FM and that they should thus be avoided.

To summarise, our results suggest that the graspability ratings are a largely valid assessment of SM despite a few items for which they remain ambiguous (see also Chapter 4 for a discussion). Action association ratings can similarly be used to reliably detect FM and non-FM objects, especially when used in conjunction with the usability judgements. Pantomime ratings closely mirror the action association ratings. As noted above, however, they should be used cautiously as they include some items that can be associated with FM on the low end of the scale. In contrast, the NoA and move-use dissimilarity instructions do not appear to bring additional information about FM. We suspect that NoA's low discriminatory power largely stems from the difficulty to determine what counts as a "typical action" with objects (e.g. do the flexion and extension of the fingers while using scissors count as 2 actions, or should *cutting* be considered a single action? Do different activities involving similar finger movements on a smartphone count as different actions?). Additionally, most manipulable objects simply tend to be associated with very few actions. Without further specifying which aspects of actions should be rated, this dimension thus does not appear to be very useful. In contrast, the move-use dissimilarity ratings likely resulted in extensive disagreements because postural differences might be too fine-grained – and their judgement too variable across participants – to be reliably assessed in a subjective norming study. It is possible that instructions closer to those used by Salmon et al. (2010) and focusing on the difference between hand *movements* instead of *postures* would result in more meaningful ratings. However, such a difference would likely be similarly difficult to judge and result in a large number of disagreements as well – as it is the case in Salmon et al.'s (2010) dataset. Other norming procedures such as asking participants to perform the two actions and then assessing their difference might yield more reliable results.

We have mentioned on several occasions that the FM ratings (usability and action association) can be used to categorise objects as FM and non-FM, thus implicitly suggesting that they should not be used as continuous variables. The first issue with treating them as continuous is that manipulability-related variables (e.g. graspability, usability and action association) are highly correlated and that they can yield unreliable results if they are included together in statistical analyses (Friedman & Wall, 2005; Hsu & Chiang, 2020; Nickels et al., 2022. See also Section 2.2.3). Second, FM-related instructions in particular tend to result in a large number of midscale items for which participants disagree. We have argued in Chapter 3 that treating such variables as linear

predictors can similarly bias the statistical estimates (Buja et al., 2019) and that non-linear models should be used instead. However, this raises another challenge regarding the interpretation of the results. Indeed, non-linear regression models do not readily allow to determine whether a predictor's effect on a dependent variable differed significantly between two portions of its range, i.e. between the low and high ends of a Likert-type scale in the current case. If a subjective variable's effect is mainly driven by the items in the middle of its scale, it could thus be difficult to conclude about any difference between items rated low and high. In light of these points, it appears more appropriate to use dichotomous categories to evaluate the effects of subjective norms, and of manipulability in particular. Control variables could arguably be used as continuous predictors, but their effects should be interpreted with caution and relative to the range of their values. Note that our recommendation goes against several authors who have argued for the use of continuous variables and regression analyses on large datasets to study the effects of various variables in word processing (e.g. Baayen, 2010; Balota et al., 2012; Brysbaert et al., 2014a, 2016). Their arguments were nevertheless mainly based on objective word characteristics (e.g. frequency of occurrence in corpora, number of letters), which are not subject to the same limitations as subjective variables – i.e. the midscale disagreement problem and validity issues. Since we have argued throughout this thesis that the latter variables cannot be considered continuous assessments of stimulus characteristics, treating them as such in statistical analyses appears difficult to defend.

5.5 Conclusion

Our analyses in the current chapter suggest that Likert-type scale ratings are best suited for general assessments of the extent to which objects are associated with structural and functional manipulations (SM and FM), but not for their finer-grained characteristics. Indeed, none of the three scales expected to capture more specific and complementary aspects of FM (number of actions, pantomime, move-use dissimilarity) were sensitive enough to their intended latent dimensions. Except for some ambiguous objects (e.g. food, clothes), the action association ratings appear to reliably capture FM on the high end of the scale. We have nevertheless also seen that this dimension led to significant disagreements overall. As argued above, it should thus not be used as a continuous variable, rather to select FM and non-FM.

General Discussion

The view that multimodal simulations mediate conceptual processing has gained considerable traction over the years. However, there also appears to be no consensus in sight given the ongoing debates and criticisms, partly because of a large number of conflicting findings and partly due to their disputed interpretations on methodological and epistemological grounds (Section 1.1.4). Focusing on manipulable objects, our own review revealed that the available evidence for the role of motor information in their representation is far from conclusive (Chapter 2). Independent of any theoretical commitment, why is it that, after roughly 25 years of research, there appears to be more controversy in this literature than any apparent progress? The goal of this thesis has been to explore one potential and largely unexplored answer to this question: the validity of our methods, and of our stimulus selection procedures in particular.

A widely used practice in experimental cognitive science is to operationally define a stimulus characteristic (e.g. manipulability) through a Likert-type scale, and to ask a group of participants to rate a list of items on the scale. The average ratings for each item are then used to select or validate the stimuli for an experiment (e.g. non-manipulable and manipulable objects). Pollock (2018) questioned this practice with one simple observation. When raters disagree in their judgements (e.g. respond on both ends of the scale), the computed average inevitably ends up in the middle of the scale – the *midscale disagreement problem*. Based on the standard deviations (SD) of the average ratings, he additionally argued that the majority of midscale items in a variety of dimensions are artefacts of the scales, not meaningful reflections of the underlying judgements. They thus do not fundamentally belong to the dimensions being assessed. As this property of ordinal scales has been largely overlooked, this observation alone raises serious concerns about the validity of a large number of findings in the literature and might explain part of the conflicting findings in the literature.

In Chapter 3, we conducted a case study on the body-object interaction (BOI, Section 2.4) ratings and provided two additional implications of the above problem. First, the presence of items for which participants do not agree entails that the ratings are potentially subject to measurement error, especially when they have been derived from a small number of observations. Although more research would be required to determine how many participants are needed, this suggests that the ratings collected in some studies might be unreliable (more on this point in Section 6.6.3).

Second, using items from the middle of the scale as stimuli introduces uncontrolled confounds that can affect an experiment's results and bias its interpretation. Indeed, something must have caused participants to disagree, and the same factors can be reasonably expected to influence task performances. A corollary of this point is that any difference in task performance observed for midscale items compared to high or low ones cannot be attributed to the dimension of interest.

The practical impact of this problem was assessed by directly inspecting the stimuli used in experiments investigating BOI's effect in word processing. We found that a large number of them unfortunately used midscale items, thus implying that their results cannot be interpreted in terms of a BOI effect. On the bright side, however, this observation also suggests that the conflicting findings in this literature (Section 2.4) can be mainly traced to methodological limitations. They thus do not represent results that must be theoretically accounted for. Although our focus in Chapter 3 was on the BOI literature, there is overall little evidence that the midscale disagreement problem has been accounted for in most studies relying on such Likert-type scale ratings for their stimuli. This suggests that there is indeed a serious and fundamental flaw in one of the most commonly used stimulus selection procedures in our discipline.

The second important contribution of Chapter 3 was the use of a new BOI dataset to understand how commonly reported summary statistics (means and SDs) relate to the amount of agreement observed among raters. This analysis provided a much clearer understanding of what these statistics truly represent and, more importantly, offered key insights for interpreting the data reported by other studies. In this sense, our results serve as a 'map' for estimating the validity of any item rated on a unipolar Likert-type scale, based on its average rating and on its SD. However, the implications of this result extend further than the interpretation of single items. We noted in Section 1.2.4 that there is, to our knowledge, no robust method for assessing the validity of subjective variables beyond inspecting their correlation with other such variables. Our newly acquired insights allow for broader investigations into the ambiguities of rating instructions and can be used to assess the overall disagreement that they generate; they thus provide a window into the validity of the instructions and of their results. We will return to this point shortly after introducing the second methodological limitation with our stimulus selection procedures explored in this thesis: operational definitions.

Subjective variables have been commonly used to assess the manipulability of objects and more generally the amount of motor information associated with them. As pointed out in Section 2.5, however, there has been little consensus on how to operationally define and assess this dimension across studies, i.e. on the instructions used to collect ratings. This point has unfortunately received little attention in the literature but is crucial for the validity and comparability of experimental findings. If two studies report conflicting results about the processing of manipulable objects but rely on different definitions of manipulability for their stimuli, they could be investigating fundamentally different constructs. Additionally, some instructions may be better suited than others to reliably capture the extent to which objects are manipulable. As there have been no comprehensive

reviews on the subject, our first goal in Chapter 4 was to identify the different instructions used for manipulability – and for structural and functional manipulability more specifically (SM and FM, Section 1.3.2) – in order to get a better sense of their variability across studies. Our review of 52 published articles revealed an unexpected degree of heterogeneity, not only in the instructions used to assess a given dimension (especially FM), but also in the interpretations of what their respective ratings fundamentally represent.

In light of this loose assessment of manipulability, our second objective in Chapter 4 was to leverage the insights gained in Chapter 3 to investigate the validity of the identified instructions. To achieve this, we partly focused on a few semantic categories representing both manipulable and non-manipulable objects (e.g. tools, food, animals, furniture, buildings) in each available dataset and inspected the rating distributions of their respective items. Following our results in Chapter 3, we divided the scales into three portions, with the middle one argued to contain items for which participants disagreed in their judgements. This approach allowed us to identify whether each category was consistently associated with low, high or midscale ratings – and thus with disagreements in the latter case. Despite some recurring disagreements, our results suggest that SM ratings are generally valid across studies and slight variations in instructions. In contrast, the assessment of FM appears to generate significant disagreements overall and to yield variable results across different rating instructions, which suggests that this dimension is particularly sensitive to subtle differences in wording. In line with our initial concerns, these observations show that the use of unstandardized instructions for manipulability – and likely for other variables – severely complicates the comparison of empirical findings in this literature, along with raising concerns about their overall reliability.

Our review in Chapter 4 also led us to the conclusion that some FM instructions target qualitatively different aspects of object use that could be theoretically relevant, while SM instructions essentially capture the same dimension. In the former case, we identified four candidate scales whose validity was difficult to fully assess because of a small number of available datasets and due to certain limitations in the wording of their instructions. These were: (i) a general assessment of the association of objects with manual use (*action association*), (ii) the number of actions typically performed with them (i.e. the diversity of uses. *Number of actions*), (iii) the ease to recognise them from a pantomime of their use (i.e. the distinctiveness of use. *Pantomime*), and (iv) the difference between the hand postures required for transport and for use (i.e. the competition between SM and FM. *Move-use dissimilarity*). The aim of Chapter 5 was to present such a multidimensional set of manipulability ratings, including one SM dimension, to both make ratings available for French words and to allow their direct comparison and assessment. Contrary to our expectations, only the SM instructions and those regarding the general association to FM yielded valid results, while the three other scales did not capture their intended construct. Additionally, the number of actions and move-use dissimilarity ratings appeared to be largely uninformative for any practical purposes. In contrast, the pantomime ratings were mostly similar to the action association ones.

The results from Chapters 4 and 5 are crucial as they highlight a deeper and more challenging issue regarding subjective variables that goes beyond the disagreement problem and the appropriate use of ratings. If the instructions do not reliably capture the dimension of interest, then the ratings become essentially meaningless and would lead to false inferences. In other words, if the low and high ends of the scale do not reflect our intended construct to begin with, identifying the items displaying a disagreement to select more appropriate stimuli would be of little use. These points have direct consequences for the validity of experiments and, if manipulability ratings are any indication, reliably capturing a stimulus characteristic through subjective scales appears to be much less straightforward than often assumed in the literature. In short, simply stating that a given dimension was assessed does not make it so; it requires further verification. Overall, the findings presented in this thesis clearly illustrate the pressing need for attention to our methods, as a lack of their validity severely undermines empirical findings and can easily lead to discrepant results that are difficult to interpret.

6.1 Open questions and limitations

The analyses conducted in this thesis are grounded in a number of assumptions and procedures that invite, and indeed warrant, further scrutiny. This section addressed some that we deem the most important.

6.1.1 The agreement metric

Most of our work relies on the level of interrater agreement, which was operationalised as the highest proportion of responses within any consecutive 3-unit interval on our 7-point scales. As discussed in Chapter 3, our main reason for using this metric was its simplicity, as it provides a straightforward indication of the aggregation in participants' ratings. Other consensus metrics that we identified in the literature (Claveria, 2021; O'Neill, 2017; Abdal Rahem & Darrah, 2018; Tastle & Wierman, 2007) were indeed either too technical to implement, difficult to interpret, or simply not adapted to the response distributions observed for Likert-type ratings[41]. Additionally, all our attempts to develop a more objective measure of agreement unfortunately yielded no satisfactory result[42].

It is important to recognise that this method remains fundamentally subjective and presents some limitations. The chosen interval length (3 units) seemed the most reasonable, as it covers

[41] For instance, entropy-based measures tend to yield high values to the items at the two ends, but also to those with strongly multimodal distributions. In contrast, items in the middle of the scale with near-Gaussian distributions receive low consensus values.

[42] Although our investigations were limited, our initial observations suggest that using a geometric approach (e.g. Claveria, 2021) or the circular mean as starting points for further development could prove promising.

conceptually close response options and allows for some variability among participants. This flexibility was particularly important for midscale items, as we observed a few with near-Gaussian distributions, albeit with naturally higher dispersion in judgements compared to items at the ends of the scale. A shorter interval would have thus overly favoured extreme items and missed many consistent responses otherwise. Conversely, a longer interval would have meant that agreement is assessed on more than half of the scale's total length, which would have undermined the metric's intent and precision. We thus believe that the 3-unit interval was the optimal choice, at least within the methodological constraints of the chosen method. The main limitation of this approach, however, is that it was specifically developed for 7-point scales and cannot be readily applied to other scale lengths. For instance, a 3-unit interval would be too wide on a 5-point scale for the same reasons highlighted above, but a 2-unit interval would be too restrictive for midscale items. We have little reason to doubt that our observations regarding the relationship between agreement levels and the summary statistics (means and standard deviations) would be fundamentally different with scales of different lengths. However, our method does not allow us to directly verify and generalise our observations, which therefore require further inquiry. A more objective and scalable consensus metric appropriate for subjective variables would be highly beneficial.

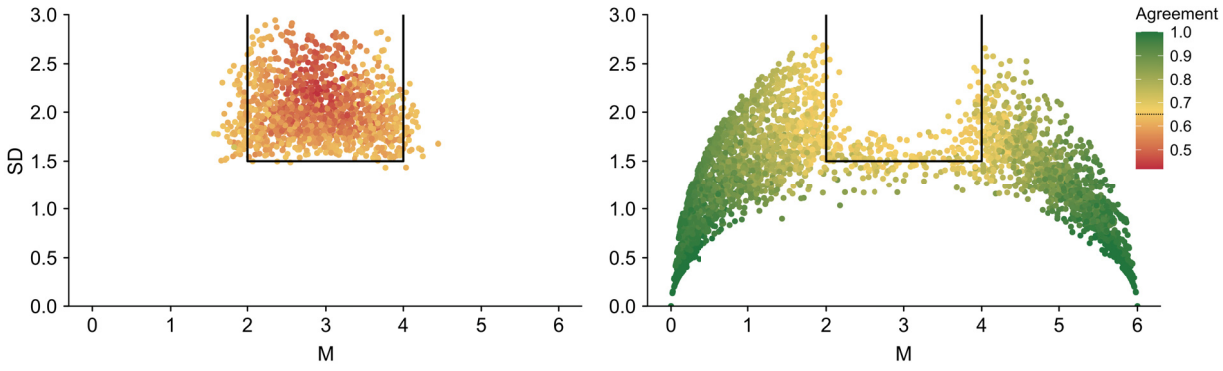
6.1.2 Our analysis of validity

As noted on several occasions, assessing the validity of subjective ratings is a daunting task as there are no available methods beyond inspecting their relationships with other variables. Chapter 4 was our attempt to approach the question through the midscale disagreement problem to gain better insights into the types of items that generate disagreements with different manipulability-related instructions. Our method proved highly informative and could be used with other dimensions to acquire a better understanding of what they capture and of their potential limitations. It should nevertheless also be pointed out that it was excessively cumbersome to implement in a review and involved some degree of speculation. Our main hypothesis starting this project was that different rating instructions would yield different results. What we did not anticipate was the significant disparity in the ratings provided by different datasets using similar instructions. Indeed, determining the extent to which the ratings were influenced by the instructions, specific methodological choices, or by the overall reliability of the datasets proved highly challenging. The issue was somewhat manageable when multiple studies used very similar instructions. In other cases, however, our conclusions necessarily relied on the ratings provided by single studies. We addressed these difficulties by conducting detailed comparative analyses between the datasets and by manually inspecting their results, which may have introduced some bias by weighing some observations more than others. We have tried to remain as cautious as possible by not reading too much into individual ratings and by aiming for more general observations. It nevertheless remains possible that some of our specific discussions were not sufficiently well-founded.

6.1.3 The area of disagreement

Our review of manipulability ratings presented in Chapter 4 primarily relied on the assumption that an average rating in the middle third of the scale and a standard deviation (SD) above half of the theoretical upper bound are symptomatic of high disagreement. Although these thresholds were coherent with our findings in Chapter 3 regarding the body-object interaction (BOI) ratings, they were also quite limited in that they were only based on a single dataset. Validating them would require further analyses that go beyond the scope of the current work. We can nevertheless make a quick verification with the additional manipulability ratings presented in Chapter 5 to ensure that our assumption holds with different instructions and more data. Figure 17 presents the identified 'area of disagreement' along with the combined unipolar Likert-type ratings collected in the current thesis (i.e. BOI, graspability, action association, pantomime, and move-use dissimilarity). Items with an agreement rate above and below .65 are further divided into two plots for better readability. Despite some overlap around the edges of the area, the plots show that the distribution of items with a low agreement rate is largely consistent with our initial assumption. Note that the .65 cut-off value is itself rather arbitrary and quite conservative. For an item with 30 observations, it would mean that only 19 participants or less rated it within 3 adjacent units of the scale. With this in mind, we are thus fairly confident about the validity of our analyses in Chapter 4.

Figure 17
Standard deviations against the means for our combined ratings, divided by low ($<$ .65, left) and high ($\geq$ .65, right) agreement items



6.1.4 Sample size considerations

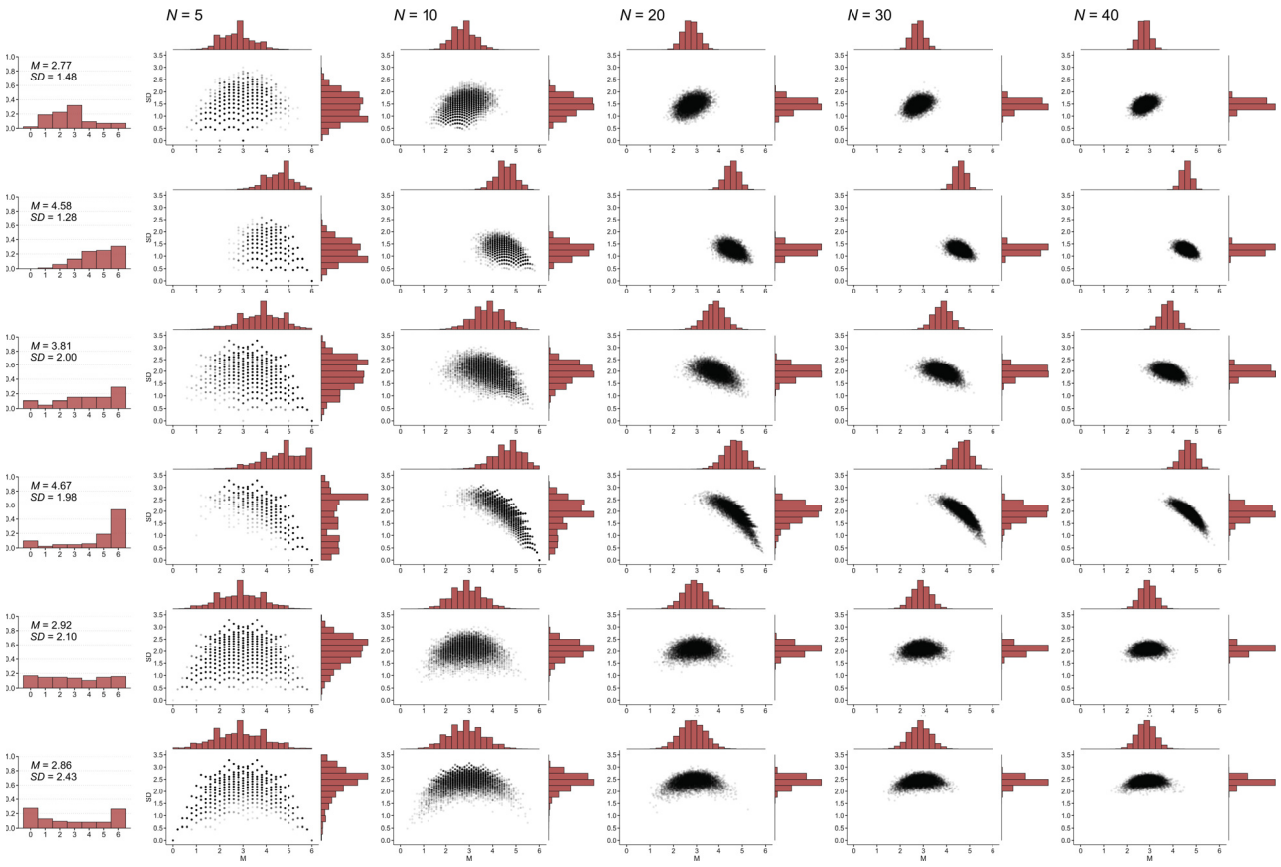
What the above discussion and our analyses in Chapter 4 did not take into account is the number of observations used to compute the ratings. We argued in Chapter 3 that one of the consequences of the disagreement problem is that it can lead to measurement error, and thus to variable summary statistics if they are not derived from a sufficient number of participants. Items that would generate a disagreement at the population level could be found outside of the area

designated above. Conversely, items with a reasonably high agreement rate in the population might end up within this zone. Thus, looking at the raw sample distributions does not inform us about where disputed or consensual items can be *expected*; it only shows where those that have been detected are. Once again, a faithful analysis of how the number of observations affects summary statistics and agreement rates would go beyond our scope. A few simple simulations based on our ratings can nevertheless at least give us a sense of the problem. The general idea behind our approach can be summarised as follows. We start with real rating distributions and treat them as what would have been observed at the population level. We then draw random samples (with replacement) from them with different sample sizes and compute their respective means and SDs. Finally, we inspect the dispersion of the summary statistics for each sample size.

Figure 18 presents such plots for 10,000 random samples of different sizes ($N = 5, 10, 20, 30, 40$), drawn from 7 distributions observed in our BOI task (Chapter 3) that did not exclusively include judgements on one side of the scale. Several interesting preliminary observations can be made on these data. First, 5 and 10 observations yield highly unstable estimates. This result is concerning because such small sample sizes are not only common in pilot rating studies aimed at stimulus control (e.g. Helbig et al., 2006; Magri et al., 2021; McNair & Harris, 2012; Rueschemeyer et al., 2010b), but also increasingly used in large-scale norming studies (approximately 10 observations by item, e.g., Muraki et al., 2022; Proos & Aigro, 2024; Winter et al., 2024). Although they retain some variability, the ratings appear to start converging at approximately 20 observations. We had argued in Chapter 3 that 30 ratings by item might not be sufficient to reliably detect the disagreements based on a comparison of different BOI datasets. The simulations nevertheless suggest that this interpretation might have been premature. Indeed, we can see that even for the items showing the largest disagreements (e.g. last row of Figure 18), the disparity in the ratings with such a sample size remains relatively constrained to the middle of the scale. Regarding our designated area of disagreement, however, there is likely a high overlap of items that have been agreed and disagreed on towards its edges, even with 40 observations (e.g. rows 1, 3, 6). For lower sample sizes, the probability of finding disputed items outside – and agreed-upon items inside – of the area increases. It is difficult to draw very clear conclusions from these simulations as they only show individual items and do not provide a bigger picture on how this variability might affect the interpretation of the ratings. We can nevertheless note that it would likely not affect our observations in Chapter 4 as they mainly relied on category-wise tendencies rather than individual item assessments.

Figure 18

Averages and standard deviations derived from random sampling (N = 10,000) across varying sample sizes and source distributions



6.2 Recommendations for unipolar Likert-type scale ratings

Researchers who want to collect or use subjective ratings are often faced with difficult choices and little methodological guidelines to inform them. The current thesis led us to a number of observations that can prove useful in this respect. It is nevertheless important to stress that some of the points discussed below were inferred from examining existing datasets and were thus not derived from systematic analyses. They should thus be treated as preliminary and hypothetical until further research is conducted to test their validity and scope.

6.2.1 Rating instructions

As previously discussed, one of the most important considerations when planning a rating study should be the wording of the instructions. This is unquestionably a highly challenging task as it is often difficult to predict their inherent ambiguities, and because we lack the proper tools to assess the validity of our constructs. It is nevertheless also evident from Chapter 4 that many researchers use ad hoc instructions, with sometimes no specific justification and despite the availability of norming studies that could inform their decisions. This practice is notably present in studies with a pilot rating phase to control their stimuli, which makes it particularly problematic because their materials are rarely shared, nor presented in much detail. Although not a solution in itself, a more careful examination of – and alignment with – the instructions already in use in the literature would facilitate the comparability of empirical findings, as well as pave the way towards better understood and more standardised assessments. More transparency is also crucial in this respect and we urge studies to make their full instructions and ratings publicly available to facilitate the evaluation of the evidence – not only their experimental results.

The validity of the instructions used by norming studies is particularly important as these often serve as reference points for future experiments and as driving forces for standardised tools. Once again, however, whether their scales faithfully capture the intended constructs is difficult to evaluate. An additional issue is that both common reliability analyses and those conducted in this thesis (Chapter 4) necessarily rely on post hoc evaluations. Collecting ratings is a tedious and costly process, especially in the current context of megastudies that include tens of thousands of entries (see Section 1.2.3). Even in light of apparent reliability and validity issues, researchers could understandably feel pressure to publish their datasets after so much effort has been dedicated to their acquisition. This raises the question of whether some steps could be implemented at earlier stages of data collection to increase data quality. Although this point requires a much wider discussion in the field, one potential solution could be to conduct smaller-scale pilot studies to allow a first assessment of the ratings. These should include a representative sample of the items planned for the main study and could be complemented with additional questions to facilitate the detection of

potential ambiguities or limitations. For instance, participants could be asked about their confidence in their responses[43], to flag ambiguous items, or more generally to provide comments if they encounter difficulty responding. Such data would in themselves be highly valuable to gain further insights into our methods and could be used to adjust the rating tasks before the main study. Note that it could also be argued that such supplementary questions should be directly integrated into rating studies. However, this can also result in much longer experimental sessions, increasing costs and the required number of participants. We thus believe that the usefulness of different alternatives should be established first.

6.2.2 Rated items and scales

Although our analyses only offer very few and speculative insights on these points, we suspect that participants do not strictly follow rating instructions. Rather, they appear to be adjusting their judgements relative to a given task's context. Two observations in particular led us to this conclusion. First, ratings appear to vary relative to the types of items included in a given study. We encountered an early example of this in Chapter 3 with the comparison of different body-object interaction datasets (BOI. Bennett et al., 2011; Pexman et al., 2019; Tillotson et al., 2008). The data provided in Pexman et al. (2019) were found to systematically have higher values than in the two other studies, with the main difference being that the former study included words from various syntactic categories (e.g. adjectives, nouns, verbs). In contrast, Bennett et al. (2011) and Tillotson et al. (2008) only normed nouns. We argued in Chapter 4 that non-noun words, which cannot be interacted with, could have been perceived as anchors on the low end of the scale, and thus biased the judgements for nouns to be higher. A more telling example is Villani et al.'s (2019) BOI norms for only abstract nouns. Given that none of their items referred to concrete objects with which it would be possible to interact with, one would expect the BOI ratings to be low. Surprisingly, almost all items had midscale or high ratings, whereas significantly lower ratings are reported for the same items in other datasets (e.g. Bennett et al., 2011; Pexman et al., 2019; Tillotson et al., 2008). It is thus likely that participants adjusted their responses relative to the items they were presented with. To ensure reliable data, these observations suggest that rating studies should provide appropriate and varied anchors, include items that could reasonably be expected on the two ends of the scale, and use only syntactic categories to which the instructions apply.

The second factor that could potentially affect the ratings is presenting participants with multiple scales for a given item. For instance, Stoinski et al. (2023) asked their participants to rate objects relative to the ease to grasp, hold and move them. The relationship of their results to related datasets suggests that the high similarity between these three dimensions led participants to focus on small differences in the instructions that were not intended by the researchers (Section 4.1.3.1). Multiple rating scales are also common in modality-specific sensory and effector-specific

[43] We would like to thank Ludovic Ferrand for suggesting this idea.

motor association norms (Section 4.3.2.2. E.g. Lynott & Connell, 2020; Repetto et al., 2022). The available data does not allow us to draw strong conclusions about how the concurrent assessment of these dimensions could have affected the results. We nevertheless suspect that participants provided contextual ratings, i.e. considering the relative association of the items to each modality or effector. It is difficult to provide clear recommendations about this point as our observations are too preliminary, but we believe that its potential effect should be at least considered in light of the dimensions being assessed.

6.2.3 Number of participants

Sample size considerations necessarily relate to the resources necessary to conduct norming studies, but can also be constrained by the limited availability of eligible participants (e.g. Proos & Aigro, 2024). Thus, collecting a too large number of ratings could quickly become excessively costly or unfeasible. Conversely, too few observations can yield highly unreliable datasets. In light of these points and of their significant implications, it is evident that the question should be investigated in much more depth to arrive at a reasonable compromise. Based on our preliminary observations in Section 6.1.4, we strongly recommend *against* collecting approximately 10 observations or less for each item. Unfortunately, the scope of the current thesis and of our analyses does not allow us to provide a precise lower bound for an adequate sample size. We would nevertheless suggest that researchers continue following the previously common practice of 30 observations per item – or more – until further studies shed more light on the question.

6.2.4 Reliability metrics

The reliability of subjective variables is directly related to the midscale disagreement problem and its statistical assessment should arguably reflect the overall disagreement present in a dataset. The three reliability metrics reported in Chapters 3 and 5 (person-total correlation, mean average absolute z score, split-half reliability[44]) were nevertheless generally difficult to interpret beyond a vague intuition about what the statistics truly imply. Split-half reliabilities were particularly insensitive to the disagreements, yielding very high values regardless of the pattern of results. The issue is that there are, to our knowledge, no available systematic analyses regarding how these metrics behave under different norming settings (e.g. randomisation procedure, number of items by participant, number of observations by item, scale length, relative proportion of low, high and midscale items). This makes their reading highly flexible and further research is needed to provide more informed guidelines. In the meantime, we recommend that rating studies provide a range of

[44]Intraclass correlations (ICC) are also commonly used by norming studies but were not suited to our experimental setup.

different reliability metrics suited to their data for transparency and to serve as a foundation for future investigations.

6.2.5 Transparency

All of the mentioned topics requiring further inquiry rely on the availability of large amounts of data, collected across different experimental settings, and that would be impossible to acquire by a single study. We would like to reiterate the critical importance of making materials and trial-level data openly available if any insights are to be gained about our tools. This would also ensure that the datasets remain useful with evolving practices and the potential introduction of new metrics.

6.2.6 Interpretation

In line with Pollock's (2018) observations, the analyses presented in Chapter 3 and the patterns observed for manipulability ratings in Chapter 4 (Section 4.2) confirm that unipolar Likert-type ratings are fundamentally closer to binary classifications than to continuous variables. This point is illustrated in Figure 19 where we have plotted the global means of all our unipolar variables as a function of their respective trimmed means for items with an agreement rate above .65. The trimmed means were calculated based on the 3-unit interval containing the highest proportion of responses for each item[45], providing a clearer indication of how the majority of participants responded. The plots show that most items received ratings clustered around the edges of the scale. For some dimensions (e.g. graspability, pantomime), there are relatively more items that appear to fall on a continuum. However, these generally remain small in number and can be difficult to distinguish from others based only on the summary statistics reported by rating studies (i.e. global means and standard deviations). Overall, we thus recommend that the default interpretation of such variables be binary. We have additionally argued throughout this thesis that the middle third portion of the scale is primarily composed of items whose average ratings are a result of a disagreement among participants. It is important to remember that this is based on a conservative threshold of agreement and that average ratings close to the middle third can also display significant disagreements – more so with smaller sample sizes.

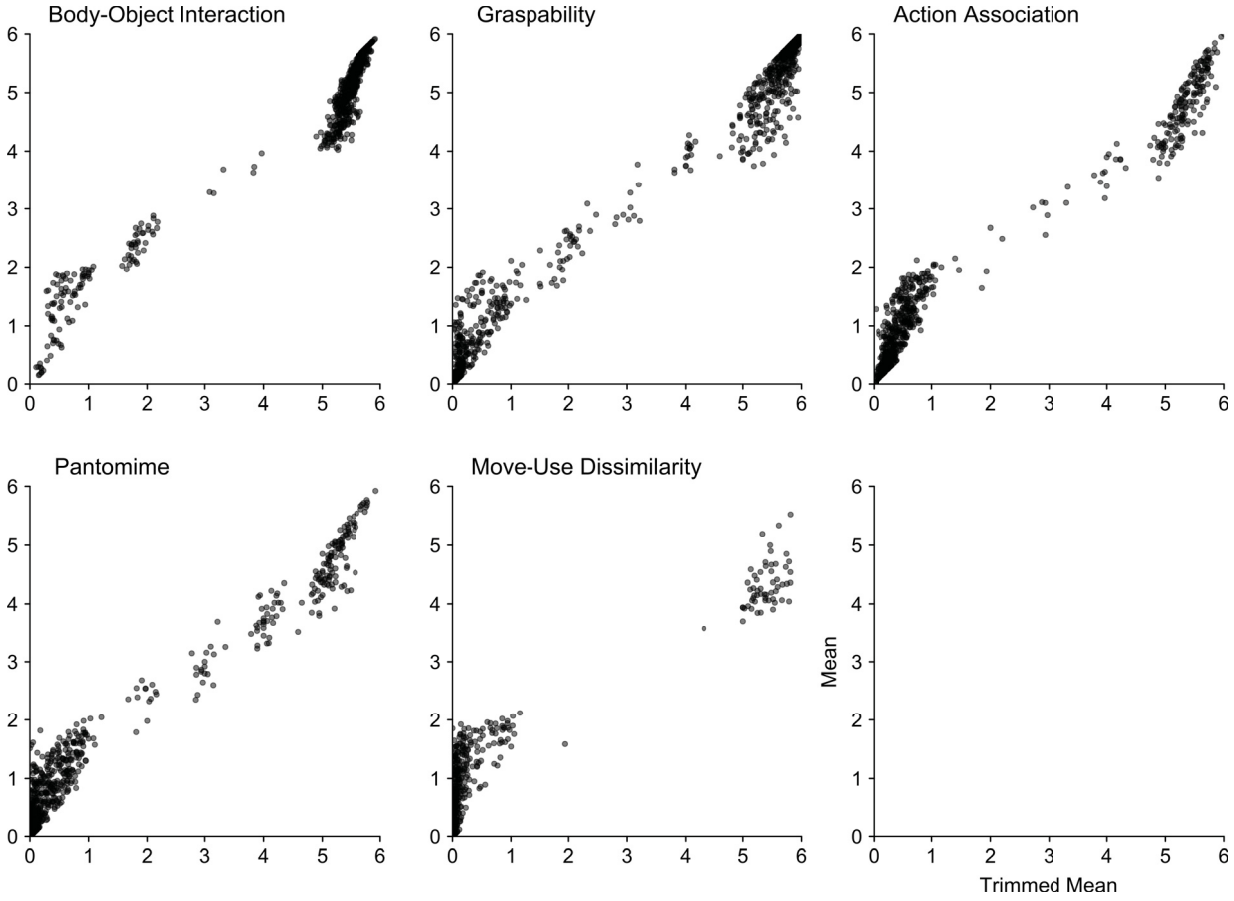
6.2.7 Use

The observation that unipolar subjective variables are binary in nature calls into question their treatment as continuous predictors in statistical models – especially in psycholinguistics where this practice is the most common (Section 1.2.3). The first issue is that the middle portion of the scale introduces potentially uncontrolled confounds due to the disagreements (Chapter 3). The nonlinear

[45]This is the same interval on which the agreement scores were computed (see Chapter 3).

Figure 19

Global means against the trimmed means for items with an agreement score above .65 across all the unipolar variables collected in the current work



relationships observed by several authors for such variables (e.g. Bonin et al., 2018; Kousta et al., 2011; Pexman et al., 2019; de Zubicaray et al., 2023) largely confirm this possibility. This, in turn, greatly complicates the interpretation of statistical results and can affect their reliability. On the one hand, using the ratings as linear predictors in the presence of a midscale effect can bias the estimates and lead to false inferences (Buja et al., 2019). On the other hand, using nonlinear models to account for such effects does not allow to easily establish statistical differences between the low and high ends of the scale (see also Chapter 5). The treatment of unipolar Likert-type ratings as continuous variables thus appears to be of little methodological or theoretical value. Based on our discussions in the previous section, we would recommend selecting items with averages as close to the ends of the scale as possible, while discarding a portion slightly larger than the middle third of the scale. Our datasets can be used to determine more specific thresholds depending on the desired level of agreement.

6.3 A measurement crisis?

This thesis unavoidably ties to broader concerns about experimental psychology's credibility, which became particularly salient with the *replicability crisis* (for discussions, see e.g., Nelson et al., 2018; Pashler & Wagenmakers, 2012). Numerous issues have been identified that pose serious threats to the reliability of published findings. Among these are notably publication biases (or the file drawer problem, Howard et al., 2009; Rosenthal, 1979), researcher degrees of freedom (Gelman & Loken, 2013; Simmons et al., 2011; Wicherts et al., 2016), questionable research practices (Banks et al., 2016; John et al., 2012), and insufficient statistical power (Ioannidis, 2005) and sampling precision (Trafimow & Myüz, 2019). Research on the role of motor representations in object processing is no exception, with most – if not all – of these problems present in the literature reviewed in Chapter 2. The field's swift response to this crisis and the ongoing changes that have ensued have nevertheless been remarkable, to say the least (see e.g., Nelson et al., 2022; Vazire, 2018). That said, to assert that "the Middle Ages are behind us, and the Enlightenment is just around the corner" (p. 529, Nelson et al., 2018) might be somewhat optimistic.

A largely underexplored question in the ever-growing meta-scientific literature has been that of measurement. Some authors have rightly stressed the foundational importance of the validity of our tools and constructs, highlighting the lack of attention they have received and the extensive methodological flexibility in current practices (e.g. Flake & Fried, 2020; Hughes, 2018; Lakens, 2024; Landy et al., 2020; Vazire et al., 2022). The results and discussions presented in the current work largely align with these concerns, revealing a general neglect of stimulus selection and norming procedures. Despite an awareness of validity issues, however, there has not been much discussion about *how* validity should be assessed. And the answer is rather simple; it is an extremely complicated epistemological problem that we do not really know how to solve in most cases. Unfortunately, ignoring it does not make it go away and, "[a]t best, we end up measuring a fuzzy version of what it is we want to study. We may even end up measuring something that isn't there at all. We say 'There it is!' while pointing vaguely in the general direction of the wrong thing" (p. 58, Hughes, 2018).

6.4 Conclusion

We began this thesis by asking whether methodological issues with the assessment and definition of stimulus characteristics could partly explain the discrepant findings on the conceptual processing of manipulable objects. In light of our investigations, the answer appears to unfortunately be a resounding 'yes'. Unfortunately – because the concerns highlighted in this thesis extend beyond the specific lines of research that we have examined, revealing deeper systemic and epistemic problems that have received little attention overall. The studies conducted in this work represent

our attempt to start addressing these issues in the context of Likert-type subjective ratings and to provide some initial guidelines for more robust methodological practices. One of our most important contributions is arguably to have mapped the midscale disagreement problem, bringing new insights into how such ratings should be interpreted and used – as well as into how disagreements can explain conflicting results. Turning to the heterogeneous definitions of manipulability, we have shown that a better understanding of the ratings' properties can be used as a window into the assessment of their validity, and ultimately to propose more informed norming instructions. Finally, our own multidimensional set of manipulability ratings has allowed us to identify the most relevant dimensions and to highlight the limitations of others that are in use in the literature. Summarising over our analyses, we have offered preliminary recommendations for the collection, interpretation and use of Likert-type ratings, and have opened several new lines of inquiry. We hope that the current work will help draw more attention to the critical importance of validating our experimental tools and that it will serve as a useful reference for future studies.

As a concluding remark, it is quite surprising that it has taken almost 70 years for someone to seriously question something as simple as what an average represents on an ordinal scale for subjective variables (Pollock, 2018). It makes one wonder if "[a]t any time the whole psychological applecart might be upset" (p. 142, Gibson, 1967).

Bibliography

Abdal Rahem, M., & Darrah, M. (2018). Using a computational approach for generalizing a consensus measure to likert scales of any size n. *International Journal of Mathematics and Mathematical Sciences*, *2018*, e5726436. https://doi.org/10.1155/2018/5726436

Adams, J. A. (1971). A closed-loop theory of motor learning. *Journal of Motor Behavior*, *3*(2), 111–150. https://doi.org/10.1080/00222895.1971.10734898

Adolph, K. E., Eppler, M. A., & Gibson, E. J. (1993). Crawling versus walking infants' perception of affordances for locomotion over sloping surfaces. *Child Development*, *64*(4), 1158–1174. https://doi.org/10.1111/j.1467-8624.1993.tb04193.x

Adolph, K. E., & Kretch, K. S. (2015, January 1). Gibson's theory of perceptual learning. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., pp. 127–134). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.23096-1

Al-Azary, H., Yu, T., & McRae, K. (2022). Can you touch the n400? the interactive effects of body-object interaction and task demands on n400 amplitudes and decision latencies. *Brain and Language*, *231*, 105147. https://doi.org/10.1016/j.bandl.2022.105147

Allport, D. A. (1985). Distributed memory, modular subsystems and dysphasia. In S. K. Newman & R. Epstein (Eds.), *Current perspectives in dysphasia* (pp. 32–60). Churchill Livingstone.

Almeida, J., Fracasso, A., Kristensen, S., Valério, D., Bergström, F., Chakravarthi, R., Tal, Z., & Walbrin, J. (2023). Neural and behavioral signatures of the multidimensionality of manipulable object processing. *Communications Biology*, *6*(1), 1–15. https://doi.org/10.1038/s42003-023-05323-x

Alonso, M. Á., Díez, E., Díez-Álamo, A. M., & Fernandez, A. (2018). Body–object interaction ratings for 750 spanish words. *Applied Psycholinguistics*, *39*(6), 1239–1252. https://doi.org/10.1017/S0142716418000309

Amsel, B. D., Urbach, T. P., & Kutas, M. (2012). Perceptual and motor attribute ratings for 559 object concepts. *Behavior Research Methods*, *44*(4), 1028–1041. https://doi.org/10.3758/s13428-012-0215-z

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355–365. https://doi.org/10.1037/0003-066X.51.4.355

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. https://doi.org/10.1037/0033-295X.111.4.1036

Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence*, *149*(1), 91–130. https://doi.org/10.1016/S0004-3702(03)00054-7

Anderson, M. L. (2017, January 17). Of Bayes and Bullets: An embodied, situated, targeting-based account of predictive processing. In T. K. Metzinger & W. Wanja (Eds.), *PPP - Philosophy and Predictive Processing*. Open MIND. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573055

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *The psychology of learning and motivation: II*, *2*, 89–195. https://doi.org/10.1016/S0079-7421(08)60422-3

Avery, J. A., Carrington, M., & Martin, A. (2023). A common neural code for representing imagined and inferred tastes. *Progress in Neurobiology*, *223*, 102423. https://doi.org/10.1016/j.pneurobio.2023.102423

Avery, J. A., Liu, A. G., Ingeholm, J. E., Riddell, C. D., Gotts, S. J., & Martin, A. (2020). Taste quality representation in the human brain. *The Journal of Neuroscience*, *40*(5), 1042–1052. https://doi.org/10.1523/JNEUROSCI.1751-19.2019

Avery, J. A., Liu, A. G., Ingeholm, J. E., Gotts, S. J., & Martin, A. (2021). Viewing images of foods evokes taste quality-specific activity in gustatory insular cortex. *Proceedings of the National Academy of Sciences*, *118*(2), e2010932118. https://doi.org/10.1073/pnas.2010932118

Azaad, S., Laham, S. M., & Shields, P. (2019). A meta-analysis of the object-based compatibility effect. *Cognition*, *190*, 105–127. https://doi.org/10.1016/j.cognition.2019.04.028

Baayen, R. H. (2010). A real experiment is a factorial experiment? *The Mental Lexicon*, *5*(1), 149–157. https://doi.org/10.1075/ml.5.1.06baa

Baddeley, A. (2018). *Exploring working memory: Selected works of alan baddeley*. Routledge. https://doi.org/10.4324/9781315111261

Balota, D. A. (1994). Visual word recognition: The journey from features to meaning. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 303–358). Academic Press.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? the role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(3), 340–357. https://doi.org/10.1037/0096-1523.10.3.340

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The english lexicon project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition* (pp. 90–115, Vol. 1). Psychology Press.

Balota, D. A., Ferraro, F. R., Connor, L. T., & Schwanenflugel, P. J. (1991). On the early influence of meaning in word recognition: A review of the literature. In *The psychology of word meanings* (pp. 199–234). Psychology Press.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*(2), 283–316. https://doi.org/10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (a travel update). In M. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 285–375). Academic Press. https://doi.org/10.1016/B978-012369374-7/50010-9

Banks, B., & Connell, L. (2022). Multi-dimensional sensorimotor grounding of concrete and abstract categories. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1870), 20210366. https://doi.org/10.1098/rstb.2021.0366

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology*, *31*(3), 323–338. https://doi.org/10.1007/s10869-016-9456-7

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*(1), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future: Topics in cognitive science. *Topics in Cognitive Science*, *2*(4), 716–724. https://doi.org/10.1111/j.1756-8765.2010.01115.x

Barsalou, L. W. (2016). On staying grounded and avoiding quixotic dead ends. *Psychonomic Bulletin & Review*, *23*(4), 1122–1142. https://doi.org/10.3758/s13423-016-1028-3

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660. https://doi.org/10.1017/S0140525X99002149

Barsalou, L. W., Kyle Simmons, W., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, *7*(2), 84–91. https://doi.org/10.1016/S1364-6613(02)00029-3

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. de Vega, A. Glenberg & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 245–284). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199217274.003.0013

Bennett, S. D. R., Burnett, A. N., Siakaluk, P. D., & Pexman, P. M. (2011). Imageability and body–object interaction ratings for 599 multisyllabic nouns. *Behavior Research Methods*, *43*(4), 1100–1109. https://doi.org/10.3758/s13428-011-0117-5

Bi, Y., Han, Z., Zhong, S., Ma, Y., Gong, G., Huang, R., Song, L., Fang, Y., He, Y., & Caramazza, A. (2015). The white matter structural network underlying human tool use and tool understanding. *The Journal of Neuroscience*, *35*(17), 6822–6835. https://doi.org/10.1523/JNEUROSCI.3709-14.2015

Bidet-Ildei, C., Meugnot, A., Beauprez, S.-A., Gimenes, M., & Toussaint, L. (2017). Short-term upper limb immobilization affects action-word understanding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1129–1139. https://doi.org/10.1037/xlm0000373

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, *33*(3), 130–174. https://doi.org/10.1080/02643294.2016.1147426

Binkofski, F., & Buxbaum, L. J. (2013). Two action systems in the human brain. *Brain and Language*, *127*(2), 222–229. https://doi.org/10.1016/j.bandl.2012.07.007

Bocanegra, Y., García, A. M., Lopera, F., Pineda, D., Baena, A., Ospina, P., Alzate, D., Buriticá, O., Moreno, L., Ibáñez, A., & Cuetos, F. (2017). Unspeakable motion: Selective action-verb impairments in parkinson's disease patients without mild cognitive impairment. *Brain and Language*, *168*, 37–46. https://doi.org/10.1016/j.bandl.2017.01.005

Bonin, P., Gelin, M., Dioux, V., & Méot, A. (2019). "it is alive!" evidence for animacy effects in semantic categorization and lexical decision. *Applied Psycholinguistics*, *40*(4), 965–985. https://doi.org/10.1017/S0142716419000092

Bonin, P., Méot, A., & Bugaiska, A. (2018). Concreteness norms for 1,659 french words: Relationships with other psycholinguistic variables and word recognition times. *Behavior Research Methods*, *50*(6), 2366–2387. https://doi.org/10.3758/s13428-018-1014-y

Bonin, P., Guillemard-Tsaparina, D., & Méot, A. (2013). Determinants of naming latencies, object comprehension times, and new norms for the russian standardized set of the colorized version of the snodgrass and vanderwart pictures. *Behavior Research Methods*, *45*(3), 731–745. https://doi.org/10.3758/s13428-012-0279-9

Bonin, P., Méot, A., Ferrand, L., & Roux, S. (2011). L'imageabilité : normes et relations avec d'autres variables psycholinguistiques: *L'Année psychologique*, *Vol. 111*(2), 327–357. https://doi.org/10.3917/anpsy.112.0327

Bonini, L., Maranesi, M., Livi, A., Fogassi, L., & Rizzolatti, G. (2014). Space-dependent representation of objects and other's action in monkey ventral premotor grasping neurons. *The Journal of Neuroscience*, *34*(11), 4108–4119. https://doi.org/10.1523/JNEUROSCI.4187-13.2014

Borghi, A. M., Bonfiglioli, C., Lugli, L., Ricciardelli, P., Rubichi, S., & Nicoletti, R. (2007). Are visual stimuli sufficient to evoke motor information? *Neuroscience Letters*, *411*(1), 17–21. https://doi.org/10.1016/j.neulet.2006.10.003

Borghi, A. M., & Riggio, L. (2015). Stable and variable affordances are both automatic and flexible. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00351

Borghi, A. M., Bonfiglioli, C., Lugli, L., Ricciardelli, P., Rubichi, S., & Nicoletti, R. (2005). Visual hand primes and manipulable objects. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *27*(27), 322–327.

Brainerd, C. J., Chang, M., Bialer, D. M., & Toglia, M. P. (2021). Semantic ambiguity and memory. *Journal of Memory and Language*, *121*, 104286. https://doi.org/10.1016/j.jml.2021.104286

Brandi, M.-L., Wohlschläger, A., Sorg, C., & Hermsdörfer, J. (2014). The neural correlates of planning and executing actual tool use. *The Journal of Neuroscience*, *34*(39), 13183–13194. https://doi.org/10.1523/JNEUROSCI.0597-14.2014

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of standardized stimuli (BOSS) phase II: 930 new normative photos (K. Paterson, Ed.). *PLoS ONE*, *9*(9), e106953. https://doi.org/10.1371/journal.pone.0106953

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The bank of standardized stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research (H. P. Op De Beeck, Ed.). *PLoS ONE*, *5*(5), e10773. https://doi.org/10.1371/journal.pone.0010773

Brodeur, M. B., Kehayia, E., Dion-Lessard, G., Chauret, M., Montreuil, T., Dionne-Dostie, E., & Lepage, M. (2012). The bank of standardized stimuli (BOSS): Comparison between french and english norms. *Behavior Research Methods*, *44*(4), 961–970. https://doi.org/10.3758/s13428-011-0184-7

Brooks, R. A. (1991). Intelligence without reason. *Proceedings of the 12th international joint conference on Artificial intelligence*, *1*, 569–595.

Brown, K. S., Yee, E., Joergensen, G., Troyer, M., Saltzman, E., Rueckl, J., Magnuson, J. S., & McRae, K. (2023). Investigating the extent to which distributional semantic models capture

a broad range of semantic relations. *Cognitive Science*, *47*(5), e13291. https://doi.org/10.1111/cogs.13291

Brunel, L., Vallet, G. T., Riou, B., Rey, A., & Versace, R. (2015). Grounded conceptual knowledge: Emergence from sensorimotor interactions. In M. H. Fischer & Y. Coello (Eds.), *Conceptual and interactive embodiment* (Vol. 2). Routledge.

Brysbaert, M., Keuleers, E., & Mandera, P. (2014a). A plea for more interactions between psycholinguistics and natural language processing research. *Computational Linguistics in the Netherlands Journal*, *4*, 209–222.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014b). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the dutch lexicon project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 441–458. https://doi.org/10.1037/xhp0000159

Brysbaert, M., Bakk, Z., Buchanan, E. M., Drieghe, D., Frey, A., Kim, E., Kuperman, V., Madan, C. R., Marelli, M., Mathôt, S., Svetina Valdivia, D., & Yap, M. (2021). Into a new decade. *Behavior Research Methods*, *53*(1), 1–3. https://doi.org/10.3758/s13428-020-01497-y

Brysbaert, M., & New, B. (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Bub, D. N., Masson, M. E., & Cree, G. S. (2008). Evocation of functional and volumetric gestural knowledge by objects and words. *Cognition*, *106*(1), 27–58. https://doi.org/10.1016/j.cognition.2006.12.010

Bub, D. N., & Masson, M. E. J. (2010). Grasping beer mugs: On the dynamics of alignment effects induced by handled objects. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(2), 341–358. https://doi.org/10.1037/a0017606

Bub, D. N., Masson, M. E. J., & Van Noordenne, M. (2021). Motor representations evoked by objects under varying action intentions. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(1), 53–80. https://doi.org/10.1037/xhp0000876

Bub, D. N., Masson, M. E. J., MacRae, C., & Marshall, G. (2018). Spatial and motor codes induced by pictures of handled objects.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, *8*(3), 531–544. https://doi.org/10.3758/BF03196189

Buhrmann, T., Di Paolo, E. A., & Barandiaran, X. (2013). A dynamical systems account of sensorimotor contingencies. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00285

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., & Zhao, L. (2019). Models as approximations i: Consequences illustrated with linear regression. *Statistical Science*, *34*(4), 523–544. https://doi.org/10.1214/18-STS693

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Buxbaum, L. J. (2001). Ideomotor apraxia: A call to action. *Neurocase*, *7*(6), 445–458. https://doi.org/10.1093/neucas/7.6.445

Buxbaum, L. J. (2017). Learning, remembering, and predicting how to use tools: Distributed neurocognitive mechanisms: Comment on osiurak and badets (2016). *Psychological Review*, *124*(3), 346–360. https://doi.org/10.1037/rev0000051

Buxbaum, L. J., & Kalénine, S. (2010). Action knowledge, visuomotor activation, and embodiment in the two action systems. *Annals of the New York Academy of Sciences*, *1191*(1), 201–218. https://doi.org/10.1111/j.1749-6632.2010.05447.x

Calzavarini, F. (2024). Rethinking modality-specificity in the cognitive neuroscience of concrete word meaning: A position paper. *Language, Cognition and Neuroscience*, *39*(7), 815–837. https://doi.org/10.1080/23273798.2023.2173789

Campanella, F., D'Agostini, S., Skrap, M., & Shallice, T. (2010). Naming manipulable objects: Anatomy of a category specific effect in left temporal tumours. *Neuropsychologia*, *48*(6), 1583–1597. https://doi.org/10.1016/j.neuropsychologia.2010.02.002

Canits, I., Pecher, D., & Zeelenberg, R. (2018). Effects of grasp compatibility on long-term memory for objects. *Acta Psychologica*, *182*, 65–74. https://doi.org/10.1016/j.actpsy.2017.11.009

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, *5*(1), 41–66. https://doi.org/10.1016/0278-2626(86)90061-8

Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience*, *10*(1), 1–34. https://doi.org/10.1162/089892998563752

Carota, F., Moseley, R., & Pulvermüller, F. (2012). Body-part-specific representations of semantic noun categories. *Journal of Cognitive Neuroscience*, *24*(6), 1492–1509. https://doi.org/10.1162/jocn_a_00219

Carroll, J. B., & White, M. N. (1973). Age-of-acquisition norms for 220 picturable nouns. *Journal of Verbal Learning and Verbal Behavior*, *12*(5), 563–576. https://doi.org/https://doi.org/10.1016/S0022-5371(73)80036-2

Casasanto, D., & Lupyan, G. (2015, May 8). All concepts are ad hoc concepts. In E. Margolis & S. Laurence (Eds.), *The conceptual mind* (pp. 543–566). The MIT Press. https://doi.org/10.7551/mitpress/9383.003.0031

Cayol, Z., & Nazir, T. A. (2020). Why language processing recruits modality specific brain regions: It is not about understanding words, but about modelling situations. *Journal of Cognition*, *3*(1), 35. https://doi.org/10.5334/joc.124

Chalard, M., Bonin, P., Méot, A., Boyer, B., & Fayol, M. (2003). Objective age-of-acquisition (AoA) norms for a set of 230 object names in french: Relationships with psycholinguistic variables, the english data from morrison et al. (1997), and naming latencies. *European Journal of Cognitive Psychology*, *15*(2), 209–245. https://doi.org/10.1080/09541440244000076

Chemero, A. (2013). Radical embodied cognitive science. *Review of General Psychology*, *17*(2), 145–150. https://doi.org/10.1037/a0032923

Chen, J., Paciocco, J. U., Deng, Z., & Culham, J. C. (2023). Human neuroimaging reveals differences in activation and connectivity between real and pantomimed tool use. *The Journal of Neuroscience*, *43*(46), 7853–7867. https://doi.org/10.1523/JNEUROSCI.0068-23.2023

Chen, S.-C., Buchanan, E. M., Kekecs, Z., Miller, J. K., Szabelska, A., Aczel, B., Bernabeu, P., Forscher, P. S., Szuts, A., Vally, D. Z., Al-Hoorie, A. H., Helmy, M., Silva, C. S. A. d., Silva, L. O. d., Moraes, Y. L. d., Hsu, R. M. C. S., Mafra, A. L., Valentova, J. V., Varella, M. A. C., … Chartier, C. R. (2024). Investigating object orientation effects across 18 languages. https://doi.org/10.31219/osf.io/2qf6w

Cho, D. T., & Proctor, R. W. (2013). Object-based correspondence effects for action-relevant and surface-property judgments with keypress responses: Evidence for a basis in spatial coding. *Psychological Research*, *77*(5), 618–636. https://doi.org/10.1007/s00426-012-0458-4

Cho, D. (, & Proctor, R. W. (2010). The object-based simon effect: Grasping affordance or relative location of the graspable part? *Journal of Experimental Psychology: Human Perception and Performance*, *36*(4), 853–861. https://doi.org/10.1037/a0019328

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, *3*(1), 1–15. https://doi.org/10.1017/S0140525X00001515

Chong, I., & Proctor, R. W. (2020). On the evolution of a radical concept: Affordances according to gibson and their subsequent use and development. *Perspectives on Psychological Science*, *15*(1), 117–132. https://doi.org/10.1177/1745691619868207

Cieslik, E. C., Zilles, K., Kurth, F., & Eickhoff, S. B. (2010). Dissociating bottom-up and top-down processes in a manual stimulus–response compatibility task. *Journal of Neurophysiology*, *104*(3), 1472–1483. https://doi.org/10.1152/jn.00261.2010

Clark, A. (1998). Embodiment and the philosophy of mind. *Royal Institute of Philosophy Supplements*, *43*, 35–51. https://doi.org/10.1017/S135824610000429X

Clark, A. (2000). *Mindware: An introduction to the philosophy of cognitive science*. Oxford University Press.

Clark, A. (2015, January 15). Predicting Peace: The End of the Representation Wars. In T. Metzinger & J. M. Windt (Eds.), *Open MIND*. Open MIND. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570979

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clarke, A. J. B., & Ludington, J. D. (2018). Thai norms for name, image, and category agreement, object familiarity, visual complexity, manipulability, and age of acquisition for 480 color photographic objects. *Journal of Psycholinguistic Research*, *47*(3), 607–626. https://doi.org/10.1007/s10936-017-9544-5

Claveria, O. (2021). A new metric of consensus for likert-type scale questionnaires: An application to consumer expectations. *Journal of Banking and Financial Technology*, *5*(1), 35–43. https://doi.org/10.1007/s42786-021-00026-5

Colling, L. J., Szucs, D., De Marco, D., Cipora, K., Ulrich, R., Nuerk, H.-C., Soltanlou, M., Bryce, D., Chen, S.-C., Schroeder, P. A., Henare, D. T., Chrystall, C. K., Corballis, P. M., Ansari, D., Goffin, C., Sokolowski, H. M., Hancock, P. J. B., Millen, A. E., Langton, S. R. H., … McShane, B. B. (2020). Registered replication report on fischer, castel, dodd, and pratt (2003). *Advances in Methods and Practices in Psychological Science*, *3*(2), 143–162. https://doi.org/10.1177/2515245920903079

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Coltheart, M. (1981). The MRC psycholinguistic database section a. *The Quarterly Journal of Experimental Psychology*, *33*(4), 497–505. https://doi.org/10.1080/14640748108400805

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Routledge.

Connell, L., Lynott, D., Fischer, M. H., & Coello, Y. (2015). Embodied semantic effects in visual word recognition. In *Conceptual and interactive embodiment* (pp. 71–92). Routledge/Taylor & Francis Group.

Connell, L., Lynott, D., & Banks, B. (2018). Interoception: The forgotten modality in perceptual grounding of abstract and concrete concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170143. https://doi.org/10.1098/rstb.2017.0143

Connell, L., & Lynott, D. (2014). Principles of representation: Why you can't represent the same concept twice. *Topics in Cognitive Science*, *6*(3), 390–406. https://doi.org/10.1111/tops.12097

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, *125*(3), 452–465. https://doi.org/10.1016/j.cognition.2012.07.010

Constant, A., Clark, A., & Friston, K. J. (2021). Representation wars: Enacting an armistice through active inference. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.598733

Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 384–387. https://doi.org/10.3758/BF03195585

Costall, A., & Morris, P. (2015). The "textbook gibson": The assimilation of dissidence. *History of Psychology*, *18*(1), 1–14. https://doi.org/10.1037/a0038398

Craighero, L., Fadiga, L., Rizzolatti, G., & Umiltà, C. (1999). Action for perception: A motor-visual attentional effect. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1673–1692. https://doi.org/10.1037/0096-1523.25.6.1673

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, *132*(2), 163–201. https://doi.org/10.1037/0096-3445.132.2.163

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we be able to run any psycholinguistic experiments at all in 1990. *Cognition*, *10*, 65–70.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*(1), 25–62. https://doi.org/10.1016/0010-0277(89)90005-X

Danker, J. F., & Anderson, J. R. (2010). The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding. *Psychological Bulletin*, *136*(1), 87–102. https://doi.org/10.1037/a0017937

Davis, C. P., Joergensen, G. H., Boddy, P., Dowling, C., & Yee, E. (2020). Making it harder to "see" meaning: The more you see something, the more its conceptual representation is susceptible to visual interference. *Psychological Science*, *31*(5), 505–517. https://doi.org/10.1177/0956797620910748

Decety, J. (1996). The neurophysiological basis of motor imagery. *Behavioural Brain Research*, *77*(1), 45–52. https://doi.org/10.1016/0166-4328(95)00225-1

Declerck, G. (2013). Why motor simulation cannot explain affordance perception. *Adaptive Behavior*, *21*(4), 286–298. https://doi.org/10.1177/1059712313488424

de Haan, E., Goodale, M. A., Jackson, S., & Schenk, T. (2018). Where to go now with "what & how". *Cortex*, *29*.

Dellantonio, S., Mulatti, C., Pastore, L., & Job, R. (2014). Measuring inconsistencies can lead you forward: Imageability and the x-ception theory. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00708

Desai, R. H., Choi, W., Lai, V. T., & Henderson, J. M. (2016). Toward semantics in the wild: Activation to manipulable nouns in naturalistic reading. *The Journal of Neuroscience*, *36*(14), 4050–4055. https://doi.org/10.1523/JNEUROSCI.1480-15.2016

Desrochers, A., & Thompson, G. L. (2009). Subjective frequency and imageability ratings for 3,600 french nouns. *Behavior Research Methods*, *41*(2), 546–557. https://doi.org/10.3758/BRM.41.2.546

de Wit, M. M., de Vries, S., van der Kamp, J., & Withagen, R. (2017). Affordances and neuroscience: Steps towards a successful marriage. *Neuroscience & Biobehavioral Reviews*, *80*, 622–629. https://doi.org/10.1016/j.neubiorev.2017.07.008

de Zubicaray, G. I., Arciuli, J., Kearney, E., Guenther, F., & McMahon, K. L. (2023). On the roles of form systematicity and sensorimotor effects in language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(3), 431–444. https://doi.org/10.1037/xlm0001201

Díez-Álamo, A. M., Díez, E., Alonso, M. Á., Vargas, C. A., & Fernandez, A. (2018). Normative ratings for perceptual and motor attributes of 750 object concepts in spanish. *Behavior Research Methods*, *50*(4), 1632–1644. https://doi.org/10.3758/s13428-017-0970-y

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research*, *91*(1), 176–180. https://doi.org/10.1007/BF00230027

Dove, G. (2011). On the need for embodied and dis-embodied cognition. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00242

Dove, G. O. (2023). Rethinking the role of language in embodied cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1870), 20210375. https://doi.org/10.1098/rstb.2021.0375

Downing-Doucet, F., & Guérard, K. (2014). A motor similarity effect in object memory. *Psychonomic Bulletin & Review*, *21*(4), 1033–1040. https://doi.org/10.3758/s13423-013-0570-5

Dreyer, F. R., Frey, D., Arana, S., Saldern, S. V., Picht, T., Vajkoczy, P., & Pulvermüller, F. (2015). Is the motor system necessary for processing action and abstract emotion words? evidence from focal brain lesions. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01661

Dreyfus, H. L. (2014, July 14). Cognitivism abandoned. In P. Baumgartner & S. Payr (Eds.), *Speaking minds: Interviews with twenty eminent cognitive scientists* (pp. 71–84). Princeton University Press. https://doi.org/10.1515/9781400863969.71

Dreyfus, H. L. (1988). The socratic and platonic basis of cognitivism. *AI and Society*, *2*(2), 99–112. https://doi.org/10.1007/bf01891374

Dreyfus, H. L. (1992). *What computers still can't do*. The MIT Press.

Duffels, B. (2022). *An examination of conceptual knowledge using near-infrared spectroscopy and electroencephalography* [Doctoral dissertation]. University of Northern British Columbia. https://doi.org/10.24124/2022/59297

Dummett, M. (1993). *The seas of language*. Clarendon Press.

Dutriaux, L., Nicolas, S., & Gyselinck, V. (2019a). Aging and posture in the memory of manipulable objects. *Aging, Neuropsychology, and Cognition*, *28*(1), 26–36. https://doi.org/10.1080/13825585.2019.1708252

Dutriaux, L., Dahiez, X., & Gyselinck, V. (2019b). How to change your memory of an object with a posture and a verb. *Quarterly Journal of Experimental Psychology*, *72*(5), 1112–1118. https://doi.org/10.1177/1747021818785096

Dutriaux, L., & Gyselinck, V. (2016). Learning is better with the hands free: The role of posture in the memory of manipulable objects (M. Iacoboni, Ed.). *PLOS ONE*, *11*(7), e0159108. https://doi.org/10.1371/journal.pone.0159108

Dutriaux, L., & Gyselinck, V. (2021). The postural effect on the memory of manipulable objects: Interference or facilitation? *Experimental Psychology*, *68*(6), 333–339. https://doi.org/10.1027/1618-3169/a000537

Eddington, C. M., & Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing: The current state of the literature. *Psychonomic Bulletin & Review*, *22*(1), 13–37. https://doi.org/10.3758/s13423-014-0665-7

Eichenbaum, H. (2010). Memory systems. *WIREs Cognitive Science*, *1*(4), 478–490. https://doi.org/10.1002/wcs.49

Ellis, R., & Tucker, M. (2000). Micro-affordance: The potentiation of components of action by seen objects. *British Journal of Psychology*, *91*(4), 451–471. https://doi.org/10.1348/000712600161934

Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (2000). Visuomotor neurons: Ambiguity of the discharge or 'motor' perception? *International Journal of Psychophysiology*, *35*(2), 165–177. https://doi.org/10.1016/S0167-8760(99)00051-3

Fagioli, S., Hommel, B., & Schubotz, R. I. (2007). Intentional control of attention: Action planning primes action-related stimulus dimensions. *Psychological Research*, *71*(1), 22–29. https://doi.org/10.1007/s00426-005-0033-3

Federico, G., Osiurak, F., Ciccarelli, G., Ilardi, C. R., Cavaliere, C., Tramontano, L., Alfano, V., Migliaccio, M., Di Cecca, A., Salvatore, M., & Brandimonte, M. A. (2023). On the functional brain networks involved in tool-related action understanding. *Communications Biology*, *6*(1), 1163. https://doi.org/10.1038/s42003-023-05518-2

Ferguson, T. D., Bub, D. N., Masson, M. E. J., & Krigolson, O. E. (2021). The role of cognitive control and top-down processes in object affordances. *Attention, Perception, & Psychophysics*, *83*(5), 2017–2032. https://doi.org/10.3758/s13414-021-02296-z

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., Conant, L. L., & Seidenberg, M. S. (2016). Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex*, *26*(5), 2018–2034. https://doi.org/10.1093/cercor/bhv020

Ferrand, L. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: Evidence from chronolex. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00306

Ferrand, L., Bonin, P., Méot, A., Augustinova, M., New, B., Pallier, C., & Brysbaert, M. (2008). Age-of-acquisition and subjective frequency estimates for all generally known monosyllabic french words and their relation with other psycholinguistic variables. *Behavior Research Methods*, *40*(4), 1049–1054. https://doi.org/10.3758/BRM.40.4.1049

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The french lexicon project: Lexical decision data for 38,840 french words and

38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488–496. https://doi.org/10.3758/BRM.42.2.488

Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., Dufau, S., Mathôt, S., & Grainger, J. (2018). MEGALEX: A megastudy of visual and auditory word recognition. *Behavior Research Methods*, *50*(3), 1285–1307. https://doi.org/10.3758/s13428-017-0943-1

Ferretti, G. (2021). A distinction concerning vision-for-action and affordance perception. *Consciousness and Cognition*, *87*, 103028. https://doi.org/10.1016/j.concog.2020.103028

Feyereisen, P., Borght, V. D., & Seron, X. (1988). The operativity effect in naming: A re-analysis. *Neuropsychologia*, *26*(3), 401–415. https://doi.org/10.1016/0028-3932(88)90094-2

Fischer, J., & Mahon, B. Z. (2021). What tool representation, intuitive physics, and action have in common: The brain's first-person physics engine. *Cognitive Neuropsychology*, *38*(7), 455–467. https://doi.org/10.1080/02643294.2022.2106126

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fodor, J. A. (1975). *The language of thought*. Crowell.

Fodor, J. A. (1983, April 6). *The modularity of mind*. MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Freud, E., Behrmann, M., & Snow, J. C. (2020). What does dorsal cortex contribute to perception? *Open Mind*, *4*, 40–56. https://doi.org/10.1162/opmi_a_00033

Friedman, L., & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, *59*(2), 127–136. https://doi.org/10.1198/000313005X41337

Gainotti, G., Spinelli, P., Scaricamazza, E., & Marra, C. (2013). The evaluation of sources of knowledge underlying different conceptual categories. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00040

Gainotti, G., Ciaraffa, F., Silveri, M. C., & Marra, C. (2009). Mental representation of normal subjects about the sources of knowledge in different semantic categories and unique entities. *Neuropsychology*, *23*(6), 803–812. https://doi.org/10.1037/a0016352

Gallese, V. (2007). Before and below 'theory of mind': Embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 659–669. https://doi.org/10.1098/rstb.2006.2002

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*(2), 593–609. https://doi.org/10.1093/brain/119.2.593

Gallivan, J. P., & Culham, J. C. (2015). Neural coding within human brain areas involved in actions. *Current Opinion in Neurobiology*, *33*, 141–149. https://doi.org/10.1016/j.conb.2015.03.012

Garcea, F. E., & Buxbaum, L. J. (2019). Gesturing tool use and tool transport actions modulates inferior parietal functional connectivity with the dorsal and ventral object processing pathways. *Human Brain Mapping*, *40*(10), 2867–2883. https://doi.org/10.1002/hbm.24565

Garcea, F. E., Chen, Q., Vargas, R., Narayan, D. A., & Mahon, B. Z. (2018). Task- and domain-specific modulation of functional connectivity in the ventral and dorsal object-processing pathways. *Brain Structure and Function*, *223*(6), 2589–2607. https://doi.org/10.1007/s00429-018-1641-1

Gardner, H. (1973). The contribution of operativity to naming capacity in aphasic patients. *Neuropsychologia*, *11*(2), 213–220. https://doi.org/https://doi.org/10.1016/0028-3932(73)90010-9

Gardner, H. (1974). The naming of objects and symbols by children and aphasic patients. *Journal of Psycholinguistic Research*, *3*(2), 133–149. https://doi.org/10.1007/BF01067572

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*(2), 256–281. https://doi.org/10.1037/0096-3445.113.2.256

Gibbs, R. W. (2006). Metaphor interpretation as embodied simulation. *Mind & Language*, *21*(3), 434–458. https://doi.org/10.1111/j.1468-0017.2006.00285.x

Gibson, E. J. (1969). *Principles of perceptual learning and development*. Appleton-Century-Crofts.

Gibson, E. J. (1988). Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. *Annual Review of Psychology*, *39*(1), 1–41.

Gibson, E. J., & Pick, A. D. (2000). *An ecological approach to perceptual learning and development*. Oxford University Press.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Gibson, J. J. (1967). James j. gibson. In *A history of psychology in autobiography, vol v.* (pp. 125–143). Appleton-Century-Crofts. https://doi.org/10.1037/11579-005

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Houghton Mifflin.

Gibson, J. J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: Towards an ecological psychology* (pp. 67–82). Lawrence Erlbaum Associates.

Gilhooly, K. J., & Logie, R. H. (1980). Meaning-dependent ratings of imagery, age of acquisition, familiarity, and concreteness for 387 ambiguous words. *Behavior Research Methods & Instrumentation*, *12*(4), 428–450. https://doi.org/10.3758/BF03201694

Gilhooly, K. J., & Watson, F. L. (1981). Word age-of-acquisition effects: A review. *Current Psychological Reviews*, *1*(3), 269–286. https://doi.org/10.1007/BF02684489

Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, *48*(3), 963–972. https://doi.org/10.3758/s13428-015-0621-0

Girardi, G., Lindemann, O., & Bekkering, H. (2010). Context effects on the processing of action-relevant object features. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(2), 330–340. https://doi.org/10.1037/a0017180

Glenberg, A. M. (2015). Few believe the world is flat: How embodiment is changing the scientific understanding of cognition. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, *69*(2), 165–171. https://doi.org/10.1037/cep0000056

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, *20*(1), 1–19. https://doi.org/10.1017/S0140525X97000010

Glenberg, A. M., Witt, J. K., & Metcalfe, J. (2013). From the revolution to embodiment: 25 years of cognitive psychology. *Perspectives on Psychological Science*, *8*(5), 573–585. https://doi.org/10.1177/1745691613498098

Glover, S. (2004). Separate visual representations in the planning and control of action. *Behavioral and Brain Sciences*, *27*(1). https://doi.org/10.1017/S0140525X04000020

Godard, M., Wamain, Y., & Kalénine, S. (2019). Do manufactured and natural objects evoke similar motor information? the case of action priming. *Quarterly Journal of Experimental Psychology*, *72*(12), 2801–2806. https://doi.org/10.1177/1747021819862210

Golonka, S., & Wilson, A. D. (2019). Ecological representations. *Ecological Psychology*, *31*(3), 235–253. https://doi.org/10.1080/10407413.2019.1615224

González, J., Barros-Loscertales, A., Pulvermüller, F., Meseguer, V., Sanjuán, A., Belloch, V., & Ávila, C. (2006). Reading cinnamon activates olfactory brain regions. *NeuroImage*, *32*(2), 906–912. https://doi.org/10.1016/j.neuroimage.2006.03.037

Goodale, M. A., Milner, A. D., Jakobson, L. S., & Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature*, *349*(6305), 154–156. https://doi.org/10.1038/349154a0

Goodale, M. A., & Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25. https://doi.org/https://doi.org/10.1016/0166-2236(92)90344-8

Grèzes, J., Tucker, M., Armony, J., Ellis, R., & Passingham, R. E. (2003). Objects automatically potentiate action: An fMRI study of implicit processing. *European Journal of Neuroscience*, *17*(12), 2735–2740. https://doi.org/10.1046/j.1460-9568.2003.02695.x

Grondin, R., Lupker, S. J., & McRae, K. (2009). Shared features dominate semantic richness effects for concrete concepts. *Journal of Memory and Language*, *60*(1), 1–19. https://doi.org/10.1016/j.jml.2008.09.001

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*(3), 377–396. https://doi.org/10.1017/S0140525X04000093

Guérard, K., Lagacé, S., & Brodeur, M. B. (2015). Four types of manipulability ratings and naming latencies for a set of 560 photographs of objects. *Behavior Research Methods*, *47*(2), 443–470. https://doi.org/10.3758/s13428-014-0488-5

Guérard, K., & Lagacé, S. (2014). A motor isolation effect: When object manipulability modulates recall performance. *Quarterly Journal of Experimental Psychology*, *67*(12), 2439–2454. https://doi.org/10.1080/17470218.2014.932399

Haddad, L., Wamain, Y., & Kalénine, S. (2024). Stimulus–response compatibility effects during object semantic categorisation: Evocation of grasp affordances or abstract coding of object size? *Quarterly Journal of Experimental Psychology*, *77*(1), 29–41. https://doi.org/10.1177/17470218231161310

Hannus, A., Cornelissen, F. W., Lindemann, O., & Bekkering, H. (2005). Selection-for-action in visual search. *Acta Psychologica*, *118*(1), 171–191. https://doi.org/https://doi.org/10.1016/j.actpsy.2004.10.010

Hansen, D., Siakaluk, P. D., & Pexman, P. M. (2012). The influence of print exposure on the body-object interaction effect in visual word recognition. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00113

Hargreaves, I. S., & Pexman, P. M. (2014). Get rich quick: The signal to respond procedure reveals the time course of semantic richness effects during visual word recognition. *Cognition*, *131*(2), 216–242. https://doi.org/10.1016/j.cognition.2014.01.001

Hargreaves, I. S., Leonard, G. A., Pexman, P. M., Pittman, D. J., Siakaluk, P. D., & Goodyear, B. G. (2012). The neural correlates of the body-object interaction effect in semantic processing. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00022

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Haro, J., & Ferré, P. (2018). Semantic ambiguity: Do multiple meanings inhibit or facilitate word recognition? *Journal of Psycholinguistic Research*, *47*(3), 679–698. https://doi.org/10.1007/s10936-017-9554-3

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*(2), 301–307. https://doi.org/10.1016/S0896-6273(03)00838-9

Heard, A., Madan, C. R., Protzner, A. B., & Pexman, P. M. (2019). Getting a grip on sensorimotor effects in lexical–semantic processing. *Behavior Research Methods*, *51*(1), 1–13. https://doi.org/10.3758/s13428-018-1072-1

Hebb, D. O. (1949). *The organization of behavior. a neuropsychological theory*. Wiley.

Helbig, H. B., Steinwender, J., Graf, M., & Kiefer, M. (2010). Action observation can prime visual object recognition. *Experimental Brain Research*, *200*(3), 251–258. https://doi.org/10.1007/s00221-009-1953-8

Helbig, H. B., Graf, M., & Kiefer, M. (2006). The role of action representations in visual object recognition. *Experimental Brain Research*, *174*(2), 221–228. https://doi.org/10.1007/s00221-006-0443-5

Heurley, L. P., Brouillet, T., Coutté, A., & Morgado, N. (2020). Size coding of alternative responses is sufficient to induce a potentiation effect with manipulable objects. *Cognition*, *205*, 104377. https://doi.org/https://doi.org/10.1016/j.cognition.2020.104377

Hoffman, P., & Lambon Ralph, M. A. (2013). Shapes, scents and sounds: Quantifying the full multi-sensory basis of conceptual knowledge. *Neuropsychologia*, *51*(1), 14–25. https://doi.org/10.1016/j.neuropsychologia.2012.11.009

Hollis, G. (2018). Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments. *Behavior Research Methods*, *50*(2), 711–729. https://doi.org/10.3758/s13428-017-0898-2

Hollis, G., & Westbury, C. (2018). When is best-worst best? a comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior Research Methods*, *50*(1), 115–133. https://doi.org/10.3758/s13428-017-1009-0

Howard, D., Best, W., Bruce, C., & Gatehouse, C. (1995). Operativity and animacy effects in aphasic naming. *International Journal of Language & Communication Disorders*, *30*(3), 286–302. https://doi.org/10.3109/13682829509021443

Hsu, N. S., Frankland, S. M., & Thompson-Schill, S. L. (2012). Chromaticity of color perception and object color knowledge. *Neuropsychologia*, *50*(2), 327–333. https://doi.org/10.1016/j.neuropsychologia.2011.12.003

Hsu, S.-Y., & Chiang, J.-T. (2022). Suppression and enhancement in multiple linear regression: A viewpoint from the perspective of a semipartial correlation coefficient. *Communications in Statistics - Theory and Methods*, *51*(7), 2057–2072. https://doi.org/10.1080/03610926.2020.1759094

Hughes, B. (2018, August 14). *Psychology in crisis*. Bloomsbury Publishing.

Hutto, D. D., & Myin, E. (2014). Neural representations not needed - no more pleas, please. *Phenomenology and the Cognitive Sciences*, *13*(2), 241–256. https://doi.org/10.1007/s11097-013-9331-1

Iachini, T., Ruggiero, G., Ruotolo, F., & Vinciguerra, M. (2014). Motor resources in peripersonal space are intrinsic to spatial encoding: Evidence from motor interference. *Acta Psychologica*, *153*, 20–27. https://doi.org/10.1016/j.actpsy.2014.09.001

Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology*, *60*, 653–670. https://doi.org/10.1146/annurev.psych.60.110707.163604

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jacob, P. (2015). Assessing radical embodiment. In Y. Coello & M. H. Fischer (Eds.), *Perceptual and emotional embodiment* (Vol. 1). Routledge.

Jakobson, L. S., Archibald, Y. M., Carey, D. P., & Goodale, M. A. (1991). A kinematic analysis of reaching and grasping movements in a patient recovering from optic ataxia. *Neuropsychologia*, *29*(8), 803–809. https://doi.org/10.1016/0028-3932(91)90073-H

James, C. T. (1975). The role of semantic information in lexical decisions. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(2), 130–136. https://doi.org/10.1037/0096-1523.1.2.130

Janyan, A., & Slavcheva, G. V. (2012). When left feels right: Asymmetry in the affordance effect. *Cognitive Processing*, *13*, 199–202. https://doi.org/10.1007/s10339-012-0450-3

Jax, S. A., & Buxbaum, L. J. (2010). Response interference between functional and structural actions linked to the same familiar object. *Cognition*, *115*(2), 350–355. https://doi.org/10.1016/j.cognition.2010.01.004

Jeannerod, M. (1986). The formation of finger grip during prehension. a cortically mediated visuomotor pattern. *Behavioural Brain Research*, *19*(2), 99–116. https://doi.org/10.1016/0166-4328(86)90008-2

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences*, *17*(2), 187–202. https://doi.org/10.1017/S0140525X00034026

Jeannerod, M. (2004). Actions from within. *International Journal of Sport and Exercise Psychology*, *2*(4), 376–402. https://doi.org/10.1080/1612197X.2004.9671752

Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage*, *14*(1), S103–S109. https://doi.org/10.1006/nimg.2001.0832

Jeannerod, M. (2003, June 6). Simulation of action as a unifying concept for motor cognition. In S. H. Johnson-Frey (Ed.), *Taking action: Cognitive neuroscience perspectives on intentional acts*. The MIT Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Johns, B. T., & Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*(1), 103–120. https://doi.org/10.1111/j.1756-8765.2011.01176.x

Johnson-Frey, S. H. (2003, June 6). Cortical representations of human tool use. In S. H. Johnson-Frey (Ed.), *Taking action: Cognitive neuroscience perspectives on intentional acts* (pp. 185–218). The MIT Press. https://doi.org/10.7551/mitpress/6614.003.0011

Johnson-Frey, S. H. (2004). The neural bases of complex tool use in humans. *Trends in Cognitive Sciences*, *8*(2), 71–78. https://doi.org/10.1016/j.tics.2003.12.002

Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition*, *13*(7), 789–845. https://doi.org/10.1080/13506280544000066

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, *131*(5), 684–712. https://doi.org/10.1037/0033-2909.131.5.684

Juhasz, B. J., Lai, Y.-H., & Woodcock, M. L. (2015). A database of 629 english compound words: Ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior Research Methods*, *47*(4), 1004–1019. https://doi.org/10.3758/s13428-014-0523-6

Juhasz, B. J., & Yap, M. J. (2013). Sensory experience ratings for over 5,000 mono- and disyllabic words. *Behavior Research Methods*, *45*(1), 160–168. https://doi.org/10.3758/s13428-012-0242-9

Kalénine, S., Buxbaum, L. J., Fischer, M. H., & Coello, Y. (2015). Role of action in conceptual object representation and organization. In *Conceptual and interactive embodiment* (Vol. 2). Routledge.

Kalénine, S., Mirman, D., Middleton, E. L., & Buxbaum, L. J. (2012). Temporal dynamics of activation of thematic and functional knowledge during conceptual processing of manipulable artifacts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1274–1295. https://doi.org/10.1037/a0027626

Kalénine, S., Shapiro, A. D., Flumini, A., Borghi, A. M., & Buxbaum, L. J. (2014). Visual context modulates potentiation of grasp types during semantic object categorization. *Psychonomic Bulletin & Review*, *21*(3), 645–651. https://doi.org/10.3758/s13423-013-0536-7

Kaschak, M. P., & Madden, J. (2021). Embodiment in the lab: Theory, measurement, and reproducibility. In M. D. Robinson & L. E. Thomas (Eds.), *Handbook of embodied psychology* (pp. 619–635). Springer International Publishing. https://doi.org/10.1007/978-3-030-78471-3_27

Kellenbach, M. L., Brett, M., & Patterson, K. (2003). Actions speak louder than functions: The importance of manipulability and action in tool representation. *Journal of Cognitive Neuroscience*, *15*(1), 30–46. https://doi.org/10.1162/089892903321107800

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample-size planning. *Evaluation & the Health Professions*, *26*(3), 258–287. https://doi.org/10.1177/0163278703255242

Kemmerer, D. (2015). Are the motor features of verb meanings represented in the precentral motor cortices? yes, but within the context of a flexible, multilevel architecture for conceptual knowledge. *Psychonomic Bulletin & Review*, *22*(4), 1068–1075. https://doi.org/10.3758/s13423-014-0784-1

Kent, C., & Lamberts, K. (2008). The encoding–retrieval relationship: Retrieval as mental simulation. *Trends in Cognitive Sciences*, *12*(3), 92–98. https://doi.org/10.1016/j.tics.2007.12.004

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behavior Research Methods*, *44*(1), 287–304. https://doi.org/10.3758/s13428-011-0118-4

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Quarterly Journal of Experimental Psychology*, *68*(8), 1457–1468. https://doi.org/10.1080/17470218.2015.1051065

Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex*, *48*(7), 805–825. https://doi.org/10.1016/j.cortex.2011.04.006

Kithu, M. C., Saccone, E. J., Crewther, S. G., Goodale, M. A., & Chouinard, P. A. (2021). A priming study on naming real versus pictures of tools. *Experimental Brain Research*, *239*(3), 821–834. https://doi.org/10.1007/s00221-020-06015-2

Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, *297*(5582), 846–848. https://doi.org/10.1126/science.1070311

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility–a model and taxonomy. *Psychological Review*, *97*(2), 253–270. https://doi.org/10.1037/0033-295X.97.2.253

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*, 14–34. https://doi.org/10.1037/a0021446

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Lachaud, C. (2007). CHACQFAM : Une base de données renseignant l'âge d'acquisition estimé et la familiarité pour 1225 mots monosyllabiques et bisyllabiques du français. *L'Année psychologique*, *107*(1), 39–63.

Lagacé, S., Downing-Doucet, F., & Guérard, K. (2013). Norms for grip agreement for 296 photographs of objects. *Behavior Research Methods*, *45*(3), 772–781. https://doi.org/10.3758/s13428-012-0283-0

Lagacé, S., & Guérard, K. (2015). When motor congruency modulates immediate memory for objects. *Acta Psychologica*, *157*, 65–73. https://doi.org/10.1016/j.actpsy.2015.02.009

Lakens, D. (2024). Concerns about replicability, theorizing, applicability, generalizability, and methodology across two crises in social psychology. https://doi.org/10.31234/osf.io/dtvs7

Lalancette, A., Garneau, É., Cochrane, A., & Wilson, M. A. (2024). Body–object interaction ratings for 3600 french nouns. *Behavior Research Methods*, *56*(7), 8009–8021. https://doi.org/10.3758/s13428-024-02466-5

Landy, J. F., Jia, M. (, Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., … Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*(5), 451–479. https://doi.org/10.1037/bul0000220

Lebois, L. A. M., Wilson-Mendenhall, C. D., & Barsalou, L. W. (2015). Are automatic conceptual cores the gold standard of semantic processing? the context-dependence of spatial meaning in grounded congruency effects. *Cognitive Science*, *39*(8), 1764–1801. https://doi.org/10.1111/cogs.12174

Lee, C.-l., Middleton, E., Mirman, D., Kalénine, S., & Buxbaum, L. J. (2013). Incidental and context-responsive activation of structure- and function-based action features during object

identification. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 257–270. https://doi.org/10.1037/a0027533

Liu, X., Banich, M. T., Jacobson, B. L., & Tanabe, J. L. (2004). Common and distinct neural substrates of attentional control in an integrated simon and spatial stroop task as assessed by event-related fMRI. *NeuroImage*, *22*(3), 1097–1106. https://doi.org/https://doi.org/10.1016/j.neuroimage.2004.02.033

Lobo, L., Heras-Escribano, M., & Travieso, D. (2018). The history and philosophy of ecological psychology. *Frontiers in Psychology*, *9*, 2228. https://doi.org/10.3389/fpsyg.2018.02228

Locher, R., Hofer, C., & Ruckstuhl, A. (2020). *IDPmisc: 'utilities of institute of data analyses and process design (www.zhaw.ch/idp)'* (Version 1.1.20).

López Zunini, R. A. (2016). *An ERP investigation of semantic richness dynamics: Multidimensionality vs. task demands* [Doctoral dissertation]. Université d'Ottawa/University of Ottawa.

Louwerse, M. M. (2008). Embodied relations are encoded in language. *Psychonomic Bulletin & Review*, *15*(4), 838–844. https://doi.org/10.3758/PBR.15.4.838

Louwerse, M. M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science*, *10*(3), 573–589. https://doi.org/10.1111/tops.12349

Lund, T. C., Sidhu, D. M., & Pexman, P. M. (2019). Sensitivity to emotion information in children's lexical processing. *Cognition*, *190*, 61–71. https://doi.org/10.1016/j.cognition.2019.04.017

Luu, D. (2017). *Keyboard latency*. Retrieved September 25, 2024, from https://danluu.com/keyboard-latency/

Lynott, D., & Connell, L. (2010). Embodied conceptual combination. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00212

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, *52*(3), 1271–1291. https://doi.org/10.3758/s13428-019-01316-z

Madan, C. R., & Singhal, A. (2012). Encoding the world around us: Motor-related processing influences verbal memory. *Consciousness and Cognition*, *21*(3), 1563–1570. https://doi.org/10.1016/j.concog.2012.07.006

Magnié, M., Besson, M., Poncet, M., & Dolisi, C. (2003). The snodgrass and vanderwart set revisited: Norms for object manipulability and for pictorial ambiguity of objects, chimeric objects, and nonobjects. *Journal of Clinical and Experimental Neuropsychology*, *25*(4), 521–560. https://doi.org/10.1076/jcen.25.4.521.13873

Magri, C., Konkle, T., & Caramazza, A. (2021). The contribution of object size, manipulability, and stability on neural responses to inanimate objects. *NeuroImage*, *237*, 118098. https://doi.org/10.1016/j.neuroimage.2021.118098

Mahon, B. Z. (2020, May 12). The representation of tools in the human brain. In D. Poeppel, G. R. Mangun & M. S. Gazzaniga (Eds.), *The cognitive neurosciences* (6th ed., pp. 765–776). The MIT Press. https://doi.org/10.7551/mitpress/11442.003.0084

Mahon, B. Z. (2015). What is embodied about cognition? *Language, Cognition and Neuroscience*, *30*(4), 420–429. https://doi.org/10.1080/23273798.2014.987791

Mahon, B. Z., Milleville, S. C., Negri, G. A., Rumiati, R. I., Caramazza, A., & Martin, A. (2007). Action-related properties shape object representations in the ventral stream. *Neuron*, *55*(3), 507–520. https://doi.org/10.1016/j.neuron.2007.07.011

Mahon, B. Z., & Hickok, G. (2016). Arguments about the nature of concepts: Symbols, embodiment, and beyond. *Psychonomic Bulletin & Review*, *23*(4), 941–958. https://doi.org/10.3758/s13423-016-1045-2

Mahon, B. Z., & Caramazza, A. (2009). Concepts and categories: A cognitive neuropsychological perspective. *Annual Review of Psychology*, *60*(1), 27–51. https://doi.org/10.1146/annurev.psych.60.110707.163532

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris*, *102*(1), 59–70. https://doi.org/10.1016/j.jphysparis.2008.03.004

Marcus-Roberts, H. M., & Roberts, F. S. (1987). Meaningless statistics. *Journal of Educational Statistics*, *12*(4), 383–394. https://doi.org/10.3102/10769986012004383

Margolis, E., & Laurence, S. (2023). Making sense of domain specificity. *Cognition*, *240*, 105583. https://doi.org/10.1016/j.cognition.2023.105583

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2015). The inter-item standard deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences*, *84*, 79–83. https://doi.org/10.1016/j.paid.2014.08.021

Marre, Q., Huet, N., & Labeye, E. (2024). Imagining abstractness: The role of embodied simulations and language in memory for abstract concepts. *Visual Cognition*, 1–24. https://doi.org/10.1080/13506285.2024.2375202

Martin, A. (2016). GRAPES—grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic Bulletin & Review*, *23*(4), 979–990. https://doi.org/10.3758/s13423-015-0842-3

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*, 25–45. https://doi.org/10.1146/annurev.psych.57.102904.190143

Masson, M. E. J. (2018). Intentions and actions. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, *72*(4), 219–228. https://doi.org/10.1037/cep0000156

Matheson, H. E., White, N. C., & McMullen, P. A. (2014a). A test of the embodied simulation theory of object perception: Potentiation of responses to artifacts and animals. *Psychological Research*, *78*(4), 465–482. https://doi.org/10.1007/s00426-013-0502-z

Matheson, H. E., White, N. C., & McMullen, P. A. (2014b). Testing the embodied account of object naming: A concurrent motor task affects naming artifacts and animals. *Acta Psychologica*, *145*, 33–43. https://doi.org/10.1016/j.actpsy.2013.10.012

Matheson, H. E., Salmon, J. P., Tougas, M., & McMullen, P. A. (2018). Embodied object concepts: The contribution of structural and functional manipulability depends on available visual information. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, *72*(4), 229–243. https://doi.org/10.1037/cep0000147

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, *27*(4), 12–12. https://doi.org/10.1609/aimag.v27i4.1904

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological Review*, *88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

McNair, N. A., & Harris, I. M. (2012). Disentangling the contributions of grasp and action representations in the recognition of manipulable objects. *Experimental Brain Research*, *220*(1), 71–77. https://doi.org/10.1007/s00221-012-3116-6

Medler, D. A. (1998). A brief history of connectionism. *Neural Computing Surveys*, *1*(2), 18–73.

Medler, D. A., Arnoldussen, A., Binder, J. R., & Seidenberg, M. S. (2005). *The wisconsin perceptual attribute ratings database*. https://www.neuro.mcw.edu/ratings/

Miklashevsky, A. (2018). Perceptual experience norms for 506 russian nouns: Modality rating, spatial localization, manipulability, imageability and other variables. *Journal of Psycholinguistic Research*, *47*(3), 641–661. https://doi.org/10.1007/s10936-017-9548-1

Miller, G. A. (1990). Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography*, *3*(4), 245–264. https://doi.org/10.1093/ijl/3.4.245

Milner, A. D., & Goodale, M. A. (1993). Visual pathways to perception and action. In T. Hicks, S. Molotchnikoff & T. Ono (Eds.). Elsevier. https://doi.org/https://doi.org/10.1016/S0079-6123(08)60379-9

Minsky, M. (1974, March 6). A framework for representing knowledge. In J. Haugeland (Ed.), *Mind design II: Philosophy, psychology, and artificial intelligence*. The MIT Press.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences*, *6*, 414–417. https://doi.org/10.1016/0166-2236(83)90190-X

Montero-Melis, G., Van Paridon, J., Ostarek, M., & Bylund, E. (2022). No evidence for embodiment: The motor system is not needed to keep action verbs in working memory. *Cortex*, *150*, 108–125. https://doi.org/10.1016/j.cortex.2022.02.006

Moreno-Martínez, F. J., Montoro, P. R., & Rodríguez-Rojo, I. C. (2014). Spanish norms for age of acquisition, concept familiarity, lexical frequency, manipulability, typicality, and other variables for 820 words from 14 living/nonliving concepts. *Behavior Research Methods*, *46*(4), 1088–1097. https://doi.org/10.3758/s13428-013-0435-x

Moreno-Martínez, F. J., & Montoro, P. R. (2012). An ecological alternative to snodgrass & vanderwart: 360 high quality colour images with norms for seven psycholinguistic variables (L. M. Martinez, Ed.). *PLoS ONE*, *7*(5), e37527. https://doi.org/10.1371/journal.pone.0037527

Moreno-Martínez, F. J., Montoro, P. R., & Laws, K. R. (2011). A set of high quality colour images with spanish norms for seven relevant psycholinguistic variables: The nombela naming test. *Aging, Neuropsychology, and Cognition*, *18*(3), 293–327. https://doi.org/10.1080/13825585.2010.540849

Moreno-Martínez, F. J., & Adrados, H. P. (2007). Un nuevo conjunto de ítems para la evaluación de la disociación ser vivo / ser no vivo con normas obtenidas de ancianos sanos españoles. *Psicológica*, *28*(1), 1–20.

Morey, R. D., Kaschak, M. P., Díez-Álamo, A. M., Glenberg, A. M., Zwaan, R. A., Lakens, D., Ibáñez, A., García, A., Gianelli, C., Jones, J. L., Madden, J., Alifano, F., Bergen, B., Bloxsom, N. G., Bub, D. N., Cai, Z. G., Chartier, C. R., Chatterjee, A., Conwell, E., … Ziv-Crispel, N. (2022). A pre-registered, multi-lab non-replication of the action-sentence compatibility effect (ACE). *Psychonomic Bulletin & Review*, *29*(2), 613–626. https://doi.org/10.3758/s13423-021-01927-8

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165–178. https://doi.org/10.1037/h0027366

Motamedi, Y., Little, H., Nielsen, A., & Sulik, J. (2019). The iconicity toolbox: Empirical approaches to measuring iconicity. *Language and Cognition*, *11*(2), 188–207. https://doi.org/10.1017/langcog.2019.14

Mountcastle, V. B., Lynch, J. C., Georgopoulos, A., Sakata, H., & Acuna, C. (1975). Posterior parietal association cortex of the monkey: Command functions for operations within extrapersonal space. *Journal of Neurophysiology*, *38*(4), 871–908. https://doi.org/10.1152/jn.1975.38.4.871

Muraki, E. J., Speed, L. J., & Pexman, P. M. (2023a). Insights into embodied cognition and mental imagery from aphantasia. *Nature Reviews Psychology*, *2*(10), 591–605. https://doi.org/10.1038/s44159-023-00221-9

Muraki, E. J., Doyle, A., Protzner, A. B., & Pexman, P. M. (2023b). Context matters: How do task demands modulate the recruitment of sensorimotor information during language processing? *Frontiers in Human Neuroscience*, *16*, 976954. https://doi.org/10.3389/fnhum.2022.976954

Muraki, E. J., Dahm, S. F., & Pexman, P. M. (2023c). Meaning in hand: Investigating shared mechanisms of motor imagery and sensorimotor simulation in language processing. *Cognition*, *240*, 105589. https://doi.org/10.1016/j.cognition.2023.105589

Muraki, E. J., Siddiqui, I. A., & Pexman, P. M. (2022). Quantifying children's sensorimotor experience: Child body–object interaction ratings for 3359 english words. *Behavior Research Methods*, *54*(6), 2864–2877. https://doi.org/10.3758/s13428-022-01798-4

Muraki, E. J., & Pexman, P. M. (2021). Simulating semantics: Are individual differences in motor imagery related to sensorimotor effects in language processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(12), 1939–1957. https://doi.org/10.1037/xlm0001039

Murata, A., Fadiga, L., Fogassi, L., Gallese, V., Raos, V., & Rizzolatti, G. (1997). Object representation in the ventral premotor cortex (area f5) of the monkey. *Journal of Neurophysiology*, *78*(4), 2226–2230. https://doi.org/10.1152/jn.1997.78.4.2226

Myung, J.-y., Blumstein, S. E., & Sedivy, J. C. (2006). Playing on the typewriter, typing on the piano: Manipulation knowledge of objects. *Cognition*, *98*(3), 223–243. https://doi.org/10.1016/j.cognition.2004.11.010

Nairne, J. S. (2002). The myth of the encoding-retrieval match. *Memory*, *10*(5), 389–395. https://doi.org/10.1080/09658210244000216

Navarrete, E., Arcara, G., Mondini, S., & Penolazzi, B. (2019). Italian norms and naming latencies for 357 high quality color images (L. Barca, Ed.). *PLOS ONE*, *14*(2), e0209524. https://doi.org/10.1371/journal.pone.0209524

Neath, I., Earle, A., Hallett, D., & Surprenant, A. M. (2011). Response time accuracy in apple macintosh computers. *Behavior Research Methods*, *43*(2), 353–362. https://doi.org/10.3758/s13428-011-0069-9

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. https://doi.org/10.3758/BF03195588

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nelson, N. C., Chung, J., Ichikawa, K., & Malik, M. M. (2022). Psychology exceptionalism and the multiple discovery of the replication crisis. *Review of General Psychology*, *26*(2), 184–198. https://doi.org/10.1177/10892680211046508

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2 : A new french lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516–524. https://doi.org/10.3758/BF03195598

New, B., ferrand, L., pallier, C., & brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the english lexicon project. *Psychonomic Bulletin & Review*, *13*(1), 45–52. https://doi.org/10.3758/BF03193811

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(4), 661–677. https://doi.org/10.1017/S014271640707035X

Newcombe, P. I., Campbell, C., Siakaluk, P. D., & Pexman, P. M. (2012). Effects of emotional and sensorimotor knowledge in semantic processing of concrete and abstract nouns. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00275

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, *4*(2), 135–183. https://doi.org/10.1016/S0364-0213(80)80015-2

Ni, L., Liu, Y., & Yu, W. (2019). The dominant role of functional action representation in object recognition. *Experimental Brain Research*, *237*(2), 363–375. https://doi.org/10.1007/s00221-018-5426-9

Nickels, L., Lampe, L. F., Mason, C., & Hameau, S. (2022). Investigating the influence of semantic factors on word retrieval: Reservations, results and recommendations. *Cognitive Neuropsychology*, *39*(3), 113–154. https://doi.org/10.1080/02643294.2022.2109958

Norman, J. (2002). Two visual systems and two theories of perception: An attempt to reconcile the constructivist and ecological approaches. *Behavioral and Brain Sciences*, *25*(1), 73–96. https://doi.org/10.1017/S0140525X0200002X

O'Neill, T. A. (2017). An overview of interrater agreement on likert scales for researchers and practitioners. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00777

Onishi, S., & Makioka, S. (2020). How a restraint of hands affects memory of hand manipulable objects: An investigation of hand position and its visibility. *Cognitive Studies: Bulletin of*

the Japanese Cognitive Science Society, *27*(3), 250–261. https://doi.org/10.11225/cs.2020.022

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973. https://doi.org/10.1017/S0140525X01000115

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*(3), 197–237. https://doi.org/10.1037/h0055737

Osiurak, F., Reynaud, E., Baumard, J., Rossetti, Y., Bartolo, A., & Lesourd, M. (2021). Pantomime of tool use: Looking beyond apraxia. *Brain Communications*, fcab263. https://doi.org/10.1093/braincomms/fcab263

Osiurak, F., Rossetti, Y., & Badets, A. (2017). What is an affordance? 40 years later. *Neuroscience & Biobehavioral Reviews*, *77*, 403–417. https://doi.org/10.1016/j.neubiorev.2017.04.014

Ostarek, M., & Huettig, F. (2019). Six challenges for embodiment research. *Current Directions in Psychological Science*, *28*(6), 593–599. https://doi.org/10.1177/0963721419866441

Ostarek, M., & Bottini, R. (2021). Towards strong inference in research on embodiment – possibilities and limitations of causal paradigms. *Journal of Cognition*, *4*(1), 5. https://doi.org/10.5334/joc.139

Oudejans, R. R. D., Michaels, C. F., Bakker, F. C., & Dolné, M. A. (1996). The relevance of action in perceiving affordances: Perception of catchableness of fly balls. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(4), 879–891. https://doi.org/10.1037/0096-1523.22.4.879

Paisios, D., Huet, N., & Labeye, E. (2023). Addressing the elephant in the middle: Implications of the midscale disagreement problem through the lens of body-object interaction ratings (D. v. Ravenzwaaij, Ed.). *Collabra: Psychology*, *9*(1), 84564. https://doi.org/10.1525/collabra.84564

Paisios, D., Labeye, E., & Huet, N. (2019). *Simulations motrices et concepts d'objets: Effets de la posture et du type de manipulabilité des objets sur leur mémoire à long terme* [Master's thesis]. Université Toulouse Jean Jaurès.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology / Revue canadienne de psychologie*, *45*(3), 255–287. https://doi.org/10.1037/h0084295

Paivio, A. (1971). Imagery and language. In S. J. Segal (Ed.), *Imagery* (pp. 7–32). Academic Press. https://doi.org/10.1016/B978-0-12-635450-8.50008-X

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, *76*(1), 1–25. https://doi.org/10.1037/h0025327

Papeo, L., Pascual-Leone, A., & Caramazza, A. (2013). Disrupting the brain to validate hypotheses on the neurobiology of language. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00148

Pappas, Z. (2014). Dissociating simon and affordance compatibility effects: Silhouettes and photographs. *Cognition*, *133*(3), 716–728. https://doi.org/https://doi.org/10.1016/j.cognition.2014.08.018

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pearson, J., & Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the National Academy of Sciences*, *112*(33), 10089–10092. https://doi.org/10.1073/pnas.1504933112

Pecher, D. (2013a). The perceptual representation of mental categories. In *The oxford handbook of cognitive psychology* (pp. 358–373). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.001.0001

Pecher, D. (2013b). No role for motor affordances in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 2–13. https://doi.org/10.1037/a0028642

Pecher, D., De Klerk, R. M., Klever, L., Post, S., Van Reenen, J. G., & Vonk, M. (2013). The role of affordances for working memory for objects. *Journal of Cognitive Psychology*, *25*(1), 107–118. https://doi.org/10.1080/20445911.2012.750324

Pecher, D., Wolters, F., & Zeelenberg, R. (2021). The role of motor action in long-term memory for objects. In M. D. Robinson & L. E. Thomas (Eds.), *Handbook of embodied psychology* (pp. 291–309). Springer International Publishing. https://doi.org/10.1007/978-3-030-78471-3_13

Pellicano, A., Iani, C., Borghi, A. M., Rubichi, S., & Nicoletti, R. (2010). Simon-like and functional affordance effects with tools: The effects of object perceptual discrimination and object action state. *Quarterly Journal of Experimental Psychology*, *63*(11), 2190–2201. https://doi.org/10.1080/17470218.2010.486903

Petilli, M. A., Günther, F., Vergallito, A., Ciapparelli, M., & Marelli, M. (2021). Data-driven computational models reveal perceptual simulation in word processing. *Journal of Memory and Language*, *117*, 104194. https://doi.org/10.1016/j.jml.2020.104194

Petrova, A., Navarrete, E., Suitner, C., Sulpizio, S., Reynolds, M., Job, R., & Peressotti, F. (2018). Spatial congruency effects exist, just not for words: Looking into estes, verges,

and barsalou (2008). *Psychological Science*, *29*(7), 1195–1199. https://doi.org/10.1177/0956797617728127

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, *15*(1), 161–167. https://doi.org/10.3758/PBR.15.1.161

Pexman, P. M. (2012). Meaning-based influences on visual word recognition. In J. S. Adelman (Ed.), *Visual word recognition: Meaning and context, individuals and development* (pp. 24–43). Psychology Press.

Pexman, P. M. (2019). The role of embodiment in conceptual development. *Language, Cognition and Neuroscience*, *34*(10), 1274–1283. https://doi.org/10.1080/23273798.2017.1303522

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The calgary semantic decision project: Concrete/abstract decision data for 10,000 english words. *Behavior Research Methods*, *49*(2), 407–417. https://doi.org/10.3758/s13428-016-0720-6

Pexman, P. M., Lupker, S. J., & Hino, Y. (2002). The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin & Review*, *9*(3), 542–549. https://doi.org/10.3758/BF03196311

Pexman, P. M., Muraki, E., Sidhu, D. M., Siakaluk, P. D., & Yap, M. J. (2019). Quantifying sensorimotor experience: Body–object interaction ratings for more than 9,000 english words. *Behavior Research Methods*, *51*(2), 453–466. https://doi.org/10.3758/s13428-018-1171-z

Phillips, C. I., Sears, C. R., & Pexman, P. M. (2012). An embodied semantic processing effect on eye gaze during sentence reading. *Language and Cognition*, *4*(2), 99–114. https://doi.org/10.1515/langcog-2012-0006

Phillips, J. C., & Ward, R. (2002). S-r correspondence effects of irrelevant visual affordance: Time course and specificity of response activation. *Visual Cognition*, *9*(4), 540–558. https://doi.org/10.1080/13506280143000575

Pisella, L., Binkofski, F., Lasek, K., Toni, I., & Rossetti, Y. (2006). No double-dissociation between optic ataxia and visual agnosia: Multiple sub-streams for multiple visuo-manual integrations. *Neuropsychologia*, *44*(13), 2734–2748. https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2006.03.027

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron*, *72*(5), 692–697. https://doi.org/10.1016/j.neuron.2011.11.001

Pollack, J. B. (1989). Connectionism: Past, present, and future. *Artificial Intelligence Review*, *3*(1), 3–20. https://doi.org/10.1007/BF00139193

Pollock, L. (2018). Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior Research Methods*, *50*(3), 1198–1216. https://doi.org/10.3758/s13428-017-0938-y

Popoviciu, T. (1965). Sur certaines inégalités qui caractérisent les fonctions convexes. *An. Stiint. Univ. "Al. I. Cuza" Iasi Sect. I a Mat.*, *11*, 155–164.

Proctor, R. W., Lien, M.-C., & Thompson, L. (2017). Do silhouettes and photographs produce fundamentally different object-based correspondence effects? *Cognition*, *169*, 91–101. https://doi.org/https://doi.org/10.1016/j.cognition.2017.08.009

Proctor, R. W., & Miles, J. D. (2014). Does the concept of affordance add anything to explanations of stimulus–response compatibility effects? In B. H. Ross (Ed.), *Psychology of learning and motivation* (pp. 227–266, Vol. 60). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-800090-8.00006-8

Proos, M., & Aigro, M. (2024). Concreteness ratings for 36,000 estonian words. *Behavior Research Methods*, *56*(5), 5178–5189. https://doi.org/10.3758/s13428-023-02257-4

Proverbio, A. M., Adorni, R., & D'Aniello, G. E. (2011). 250ms to code for action affordance during observation of manipulable objects. *Neuropsychologia*, *49*(9), 2711–2717. https://doi.org/10.1016/j.neuropsychologia.2011.05.019

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, *22*(2), 253–279. https://doi.org/10.1017/S0140525X9900182X

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin*, *80*(1), 1–24. https://doi.org/10.1037/h0034650

Quak, M., Pecher, D., & Zeelenberg, R. (2014). Effects of motor congruence on visual working memory. *Attention, Perception, & Psychophysics*, *76*(7), 2063–2070. https://doi.org/10.3758/s13414-014-0654-y

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, *12*(5), 410–430. https://doi.org/10.1002/bs.3830120511

Raos, V., Umiltá, M.-A., Murata, A., Fogassi, L., & Gallese, V. (2006). Functional properties of grasping-related neurons in the ventral premotor area f5 of the macaque monkey. *Journal of Neurophysiology*, *95*(2), 709–729. https://doi.org/10.1152/jn.00463.2005

Rastle, K. (2016, January 1). Visual word recognition. In *Neurobiology of language* (pp. 255–264). Academic Press. https://doi.org/10.1016/B978-0-12-407794-2.00021-3

Repetto, C., Rodella, C., Conca, F., Santi, G. C., & Catricalà, E. (2023). The italian sensorimotor norms: Perception and action strength measures for 959 words. *Behavior Research Methods*, *55*(8), 4035–4047. https://doi.org/10.3758/s13428-022-02004-1

Righi, S., Orlando, V., & Marzi, T. (2014). Attractiveness and affordance shape tools neural coding: Insight from ERPs. *International Journal of Psychophysiology*, *91*(3), 240–253. https://doi.org/10.1016/j.ijpsycho.2014.01.003

Rio, L. (2021, October 18). *Affordances and language: How the level of object familiarity modulates the manipulation and categorization of objects across development* [Doctoral dissertation]. Alma Mater Studiorum - Università di Bologna. https://doi.org/10.48676/unibo/amsdottorato/9959

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, *12*(1), 1–20. https://doi.org/10.1016/S0022-5371(73)80056-8

Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., & Matelli, M. (1988). Functional organization of inferior area 6 in the macaque monkey. *Experimental Brain Research*, *71*(3), 491–507. https://doi.org/10.1007/BF00248742

Rizzolatti, G., & Fadiga, L. (1998). Grasping objects and grasping action meanings: The dual role of monkey rostroventral premotor cortex (area f5). In G. R. Bock & J. A. Goode (Eds.), *Sensory guidance of movement* (pp. 81–95). John Wiley & Sons.

Rizzolatti, G., & Matelli, M. (2003). Two different streams form the dorsal visual system: Anatomy and functions. *Experimental Brain Research*, *153*(2), 146–157. https://doi.org/10.1007/s00221-003-1588-0

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, *10*(1), 18–24. https://doi.org/10.1086/286788

Rossetti, Y., Pisella, L., & McIntosh, R. D. (2017). Rise and fall of the two visual systems theory. *Annals of Physical and Rehabilitation Medicine*, *60*(3), 130–140. https://doi.org/10.1016/j.rehab.2017.02.002

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 487–494. https://doi.org/10.1016/S0022-5371(70)80091-3

Rueschemeyer, S.-A., Lindemann, O., Van Rooij, D., Van Dam, W., & Bekkering, H. (2010a). Effects of intentional motor actions on embodied language processing. *Experimental Psychology*, *57*(4), 260–266. https://doi.org/10.1027/1618-3169/a000031

Rueschemeyer, S.-A., Van Rooij, D., Lindemann, O., Willems, R. M., & Bekkering, H. (2010b). The function of words: Distinct neural correlates for words denoting differently manipulable objects. *Journal of Cognitive Neuroscience*, *22*(8), 1844–1851. https://doi.org/10.1162/jocn.2009.21310

Rumelhart, D. E., McClelland, J. L., & Group, P. R. (1986, July 17). *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT Press. https://doi.org/10.7551/mitpress/5236.001.0001

Ruotolo, F., Kalénine, S., & Bartolo, A. (2020). Activation of manipulation and function knowledge during visual search for objects. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(1), 66–90. https://doi.org/10.1037/xhp0000696

Ruspoli, T. (2010). *Being in the world*.

Saccone, E. J., Churches, O., & Nicholls, M. E. R. (2016). Explicit spatial compatibility is not critical to the object handle effect. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(10), 1643–1653. https://doi.org/10.1037/xhp0000258

Saccone, E. J., Thomas, N. A., & Nicholls, M. E. R. (2021). One-handed motor activity does not interfere with naming lateralized pictures of tools. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(4), 529–544. https://doi.org/10.1037/xhp0000863

Sakreida, K., Effnert, I., Thill, S., Menz, M. M., Jirak, D., Eickhoff, C. R., Ziemke, T., Eickhoff, S. B., Borghi, A. M., & Binkofski, F. (2016). Affordance processing in segregated parieto-frontal dorsal stream sub-pathways. *Neuroscience & Biobehavioral Reviews*, *69*, 89–112. https://doi.org/10.1016/j.neubiorev.2016.07.032

Salmon, J. P., McMullen, P. A., & Filliter, J. H. (2010). Norms for two types of manipulability (graspability and functional usage), familiarity, and age of acquisition for 320 photographs of objects. *Behavior Research Methods*, *42*(1), 82–95. https://doi.org/10.3758/BRM.42.1.82

Salmon, J. P., E., H., & McMullen, P. A. (2014). Photographs of manipulable objects are named more quickly than the same objects depicted as line-drawings: Evidence that photographs engage embodiment more than line-drawings. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01187

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*, *8*(9), 657–661. https://doi.org/10.1038/nrn2213

Schank, R. C., & Abelson, R. P. (1975). Scripts, plans, and knowledge. *Proceedings of the 4th international joint conference on Artificial intelligence - Volume 1*, *1*, 151–157.

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007467. https://doi.org/10.1177/25152459211007467

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review*, *82*(4), 225–260. https://doi.org/10.1037/h0076770

Schneider, G. E. (1969). Two visual systems. *Science*, *163*(3870), 895–902. https://doi.org/10.1126/science.163.3870.895

Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, *3*(3), 417–457.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568. https://doi.org/10.1037/0033-295X.96.4.523

Serious Science. (2017, June 16). *Free energy principle — karl friston*. Retrieved October 4, 2024, from https://www.youtube.com/watch?v=NIu_dJGyIQI

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press. https://doi.org/10.1017/CBO9780511526817

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shapiro, L. A. (2011). Embodied cognition: Lessons from linguistic determinism. *Philosophical Topics*, *39*(1), 121–140.

Shin, Y. K., Proctor, R. W., & Capaldi, E. J. (2010). A review of contemporary ideomotor theory. *Psychological Bulletin*, *136*(6), 943–974. https://doi.org/10.1037/a0020541

Siakaluk, P. D., Pexman, P. M., Aguilera, L., Owen, W. J., & Sears, C. R. (2008a). Evidence for the activation of sensorimotor information during visual word recognition: The body–object interaction effect. *Cognition*, *106*(1), 433–443. https://doi.org/10.1016/j.cognition.2006.12.011

Siakaluk, P. D., Pexman, P. M., Sears, C. R., Wilson, K., Locheed, K., & Owen, W. J. (2008b). The benefits of sensorimotor knowledge: Body–object interaction facilitates semantic processing. *Cognitive Science*, *32*(3), 591–605. https://doi.org/10.1080/03640210802035399

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, W. K., Rapuano, K. M., Kallman, S. J., Ingeholm, J. E., Miller, B., Gotts, S. J., Avery, J. A., Hall, K. D., & Martin, A. (2013). Category-specific integration of homeostatic signals in caudal but not rostral human insula. *Nature Neuroscience*, *16*(11), 1551–1552. https://doi.org/10.1038/nn.3535

Simmons, W. K., Ramjee, V., Beauchamp, M. S., McRae, K., Martin, A., & Barsalou, L. W. (2007). A common neural substrate for perceiving and knowing about color. *Neuropsychologia*, *45*(12), 2802–2810. https://doi.org/10.1016/j.neuropsychologia.2007.05.002

Simon, J. R. (1990). The effects of an irrelevant directional CUE on human information processing. In R. W. Proctor & T. G. Reeve (Eds.), *Stimulus-response compatibility* (pp. 31–86, Vol. 65). North-Holland. https://doi.org/https://doi.org/10.1016/S0166-4115(08)61218-2

Simon, J. R. (1969). Reactions toward the source of stimulation. *Journal of Experimental Psychology*, *81*(1), 174–176. https://doi.org/10.1037/h0027448

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*(3), 214–241. https://doi.org/10.1037/h0036351

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–23. https://doi.org/10.1017/S0140525X00052432

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Solana, P., & Santiago, J. (2022). Does the involvement of motor cortex in embodied language comprehension stand on solid ground? a p-curve analysis and test for excess significance of the TMS and tDCS evidence. *Neuroscience & Biobehavioral Reviews*, *141*, 104834. https://doi.org/10.1016/j.neubiorev.2022.104834

Solana, P., & Santiago, J. (2023, March 31). Worse than expected: A z-curve reanalysis of motor cortex stimulation studies of embodied language comprehension. https://doi.org/10.31234/osf.io/7sptf

Spreen, O., & Schulz, R. W. (1966). Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior*, *5*(5), 459–468. https://doi.org/10.1016/S0022-5371(66)80061-0

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177. https://doi.org/10.1016/j.nlm.2004.06.005

Stanners, R. F., Jastrzembski, J. E., & Westbrook, A. (1975). Frequency and visual quality in a word-nonword classification task. *Journal of Verbal Learning and Verbal Behavior*, *14*(3), 259–264. https://doi.org/10.1016/S0022-5371(75)80069-7

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78. https://doi.org/10.1207/s15516709cog2901_3

Stoinski, L. M., Perkuhn, J., & Hebart, M. N. (2023). THINGSplus: New norms and metadata for the THINGS database of 1854 object concepts and 26,107 natural object images. *Behavior Research Methods*, *56*(3), 1583–1603. https://doi.org/10.3758/s13428-023-02110-8

Strik Lievers, F., Bolognesi, M., & Winter, B. (2021). The linguistic dimensions of concrete and abstract concepts: Lexical category, morphological structure, countability, and etymology. *Cognitive Linguistics*, *32*(4), 641–670. https://doi.org/10.1515/cog-2021-0007

Suárez-García, D. M. A., Birba, A., Zimerman, M., Diazgranados, J. A., Lopes Da Cunha, P., Ibáñez, A., Grisales-Cárdenas, J. S., Cardona, J. F., & García, A. M. (2021). Rekindling action language: A neuromodulatory study on parkinson's disease patients. *Brain Sciences*, *11*(7), 887. https://doi.org/10.3390/brainsci11070887

Suzuki, T., Takagi, M., & Sugawara, K. (2012). Affordance effects in grasping actions for graspable objects: Electromyographic reaction time study. *Perceptual and Motor Skills*, *115*(3), 881–890. https://doi.org/10.2466/26.22.24.PMS.115.6.881-890

Symes, E., Ellis, R., & Tucker, M. (2005). Dissociating object-based and space-based affordances. *Visual Cognition*, *12*(7), 1337–1361. https://doi.org/10.1080/13506280444000445

Sze, W. P., Rickard Liow, S. J., & Yap, M. J. (2014). The chinese lexicon project: A repository of lexical decision behavioral responses for 2,500 chinese characters. *Behavior Research Methods*, *46*(1), 263–273. https://doi.org/10.3758/s13428-013-0355-9

Taikh, A., Hargreaves, I. S., Yap, M. J., & Pexman, P. M. (2015). Semantic classification of pictures and words. *Quarterly Journal of Experimental Psychology*, *68*(8), 1502–1518. https://doi.org/10.1080/17470218.2014.975728

Taira, M., Mine, S., Georgopoulos, A., Murata, A., & Sakata, H. (1990). Parietal cortex neurons of the monkey related to the visual guidance of hand movement. *Experimental Brain Research*, *83*(1). https://doi.org/10.1007/BF00232190

Tanné-Gariépy, J., Rouiller, E. M., & Boussaoud, D. (2002). Parietal inputs to dorsal versus ventral premotor areas in the macaque monkey: Evidence for largely segregated visuomotor pathways. *Experimental Brain Research*, *145*(1), 91–103. https://doi.org/10.1007/s00221-002-1078-9

Tastle, W. J., & Wierman, M. J. (2007). Consensus and dissention: A measure of ordinal dispersion. *International Journal of Approximate Reasoning*, *45*(3), 531–545. https://doi.org/10.1016/j.ijar.2006.06.024

Taylor, J. E., Rousselet, G. A., Scheepers, C., & Sereno, S. C. (2022). Rating norms should be calculated from cumulative link mixed effects models. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01814-7

Team, R. C. (2021). *R: A language and environment for statistical computing* (Version 4.1.2). R Foundation for Statistical Computing.

Team, R. C. (2023). *R: A language and environment for statistical computing* (Version 4.3.2). R Foundation for Statistical Computing.

Team, R. (2022). *RStudio: Integrated development environment for r* (Version 2022.7.0.548).

Team, R. (2024). *RStudio: Integrated development environment for r* (Version 2023.12.1.402).

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S. F., & Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, *17*(2), 273–281. https://doi.org/10.1162/0898929053124965

Tettamanti, M., Conca, F., Falini, A., & Perani, D. (2017). Unaware processing of tools in the neural system for object-directed action representation. *The Journal of Neuroscience*, *37*(44), 10712–10724. https://doi.org/10.1523/JNEUROSCI.1061-17.2017

Thomas, E. R., Stötefalk, N., Pecher, D., & Zeelenberg, R. (2019). Alignment effects for pictured objects: Do instructions to "imagine picking up an object" prime actions? *Journal of Experimental Psychology: Human Perception and Performance*, *45*(10), 1346–1354. https://doi.org/10.1037/xhp0000676

Tillotson, S. M., Siakaluk, P. D., & Pexman, P. M. (2008). Body—object interaction ratings for 1,618 monosyllabic nouns. *Behavior Research Methods*, *40*(4), 1075–1078. https://doi.org/10.3758/BRM.40.4.1075

Tipper, S. P., Paul, M. A., & Hayes, A. E. (2006). Vision-for-action: The effects of object property discrimination and action state on affordance compatibility effects. *Psychonomic Bulletin & Review*, *13*(3), 493–498. https://doi.org/10.3758/BF03193875

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Lawrence Erlbaum.

Tomassini, V., Jbabdi, S., Klein, J. C., Behrens, T. E. J., Pozzilli, C., Matthews, P. M., Rushworth, M. F. S., & Johansen-Berg, H. (2007). Diffusion-weighted imaging tractography-based parcellation of the human lateral premotor cortex identifies dorsal and ventral subregions with anatomical and functional specializations. *The Journal of Neuroscience*, *27*(38), 10259–10269. https://doi.org/10.1523/JNEUROSCI.2144-07.2007

Tousignant, C., & Pexman, P. M. (2012). Flexible recruitment of semantic richness: Context modulates body-object interaction effects in lexical-semantic processing. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00053

Toussaint, L., Wamain, Y., Bidet-Ildei, C., & Coello, Y. (2020). Short-term upper-limb immobilization alters peripersonal space representation. *Psychological Research*, *84*(4), 907–914. https://doi.org/10.1007/s00426-018-1118-0

Trafimow, D. (2018). An a priori solution to the replication crisis. *Philosophical Psychology*, *31*(8), 1188–1214. https://doi.org/10.1080/09515089.2018.1490707

Trafimow, D., & Myüz, H. A. (2019). The sampling precision of research in five major areas of psychology. *Behavior Research Methods*, *51*(5), 2039–2058. https://doi.org/10.3758/s13428-018-1173-x

Tranel, D., Logan, C. G., Frank, R. J., & Damasio, A. R. (1997). Explaining category-related effects in the retrieval of conceptual and lexical knowledge for concrete entities: Operationalization and analysis of factors. *Neuropsychologia*, *35*(10), 1329–1339. https://doi.org/10.1016/S0028-3932(97)00086-9

Tucker, M., & Ellis, R. (2004). Action priming by briefly presented objects. *Acta Psychologica*, *116*(2), 185–203. https://doi.org/https://doi.org/10.1016/j.actpsy.2004.01.004

Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 830–846. https://doi.org/10.1037/0096-1523.24.3.830

Tucker, M., & Ellis, R. (2001). The potentiation of grasp types during visual object categorization. *Visual Cognition*, *8*(6), 769–800. https://doi.org/10.1080/13506280042000144

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory*. Academic Press.

Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*, 230–265. https://doi.org/10.1112/plms/s2-42.1.230

Turvey, M. T., Fitch, H. L., Tuller, B., & Kelso, J. A. S. (1982). The bernstein perspective: I. the problems of degrees of freedom and context-conditioned variability. In *Human motor behavior* (pp. 239–252). Psychology Press.

Ungerleider, M., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale & R. J. Mansfield (Eds.), *Analysis of visual behavior*. The MIT Press.

Vainio, L., Symes, E., Ellis, R., Tucker, M., & Ottoboni, G. (2008). On the relations between action planning, object identification, and motor representations of observed actions and objects. *Cognition*, *108*(2), 444–465. https://doi.org/10.1016/j.cognition.2008.03.007

Valyear, K. F., Chapman, C. S., Gallivan, J. P., Mark, R. S., & Culham, J. C. (2011). To use or to move: Goal-set modulates priming when grasping real tools. *Experimental Brain Research*, *212*(1), 125–142. https://doi.org/10.1007/s00221-011-2705-0

VanArsdall, J. E., & Blunt, J. R. (2022). Analyzing the structure of animacy: Exploring relationships among six new animacy and 15 existing normative dimensions for 1,200 concrete

nouns. *Memory & Cognition*, *50*(5), 997–1012. https://doi.org/10.3758/s13421-021-01266-y

van Elk, M., Van Schie, H., & Bekkering, H. (2014). Action semantics: A unifying conceptual framework for the selective use of multimodal and modality-specific object knowledge. *Physics of Life Reviews*, *11*(2), 220–250. https://doi.org/10.1016/j.plrev.2013.11.005

van Hoef, R., Connell, L., & Lynott, D. (2023). The effects of sensorimotor and linguistic information on the basic-level advantage. *Cognition*, *241*, 105606. https://doi.org/10.1016/j.cognition.2023.105606

Varela, F. J., Rosch, E., & Thompson, E. (1991, September 26). *The embodied mind: Cognitive science and human experience*. The MIT Press. https://doi.org/10.7551/mitpress/6730.001.0001

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, *31*(2), 162–168. https://doi.org/10.1177/09637214211067779

Versace, R., Labeye, É., Badard, G., & Rose, M. (2009). The contents of long-term memory and the emergence of knowledge. *European Journal of Cognitive Psychology*, *21*(4), 522–560. https://doi.org/10.1080/09541440801951844

Villani, C., Lugli, L., Liuzza, M. T., & Borghi, A. M. (2019). Varieties of abstract concepts and their multiple dimensions. *Language and Cognition*, *11*(3), 403–430. https://doi.org/10.1017/langcog.2019.23

Vingerhoets, G., Vandamme, K., & Vercammen, A. (2009). Conceptual and physical object qualities contribute differently to motor affordances. *Brain and Cognition*, *69*(3), 481–489. https://doi.org/10.1016/j.bandc.2008.10.003

Vogt, S., Taylor, P., & Hopkins, B. (2003). Visuomotor priming by pictures of hand postures: Perspective matters. *Neuropsychologia*, *41*(8), 941–951. https://doi.org/https://doi.org/10.1016/S0028-3932(02)00319-6

von Holst, E. (1954). Relations between the central nervous system and the peripheral organs. *British Journal of Animal Behaviour*, *2*, 89–94. https://doi.org/10.1016/S0950-5601(54)80044-X

von Neumann, J. (1993). First draft of a report on the EDVAC. *IEEE Annals of the History of Computing*, *15*(4), 27–75. https://doi.org/10.1109/85.238389

von Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschung*, *18*(1), 299–342. https://doi.org/10.1007/BF02409636

Wamain, Y., Sahaï, A., Decroix, J., Coello, Y., & Kalénine, S. (2018). Conflict between gesture representations extinguishes mu rhythm desynchronization during manipulable object perception: An EEG study. *Biological Psychology*, *132*, 202–211. https://doi.org/https://doi.org/10.1016/j.biopsycho.2017.12.004

Wang, S., Zhang, Y., Shi, W., Zhang, G., Zhang, J., Lin, N., & Zong, C. (2023). A large dataset of semantic ratings and its computational extension. *Scientific Data*, *10*(1), 106. https://doi.org/10.1038/s41597-023-01995-6

Warren, W. H. (1984). Perceiving affordances: Visual guidance of stair climbing. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(5), 683–703. https://doi.org/10.1037/0096-1523.10.5.683

Warrington, E. K., & McCarthy, R. A. (1987). Categories of knowledge: Further fractionations and an attempted integration. *Brain*, *110*(5), 1273–1296. https://doi.org/10.1093/brain/110.5.1273

Warrington, E. K., & McCarthy, R. A. (1983). Category specific access dysphasia. *Brain*, *106*(4), 859–878. https://doi.org/10.1093/brain/106.4.859

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*(3), 829–853. https://doi.org/10.1093/brain/107.3.829

Wellsby, M., Siakaluk, P. D., Owen, W. J., & Pexman, P. M. (2011). Embodied semantic processing: The body-object interaction effect in a non-manual task. *Language and Cognition*, *3*(1), 1–14. https://doi.org/10.1515/langcog.2011.001

Wellsby, M., & Pexman, P. M. (2014). The influence of bodily experience on children's language processing. *Topics in Cognitive Science*, *6*(3), 425–441. https://doi.org/10.1111/tops.12092

Whaley, C. (1978). Word—nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, *17*(2), 143–154. https://doi.org/10.1016/S0022-5371(78)90110-X

Wheeler, M. (2005). *Reconstructing the cognitive world: The next step*. MIT Press.

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Van Aert, R. C. M., & Van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636. https://doi.org/10.3758/BF03196322

Wingfield, C., & Connell, L. (2022). Understanding the role of linguistic distributional knowledge in cognition. *Language, Cognition and Neuroscience*, *37*(10), 1220–1270. https://doi.org/10.1080/23273798.2022.2069278

Winter, B., Lupyan, G., Perry, L. K., Dingemanse, M., & Perlman, M. (2024). Iconicity ratings for 14,000+ english words. *Behavior Research Methods*, *56*(3), 1640–1655. https://doi.org/10.3758/s13428-023-02112-6

Witt, J. K., Kemmerer, D., Linkenauger, S. A., & Culham, J. (2010). A functional role for motor simulation in identifying tools. *Psychological Science*, *21*(9), 1215–1219. https://doi.org/10.1177/0956797610378307

Witt, J. K., Kemmerer, D., Linkenauger, S. A., & Culham, J. C. (2020). Reanalysis suggests evidence for motor simulation in naming tools is limited: A commentary on witt, kemmerer, linkenauger, and culham (2010). *Psychological Science*, *31*(8), 1036–1039. https://doi.org/10.1177/0956797620940555

Wykowska, A., Schubö, A., & Hommel, B. (2009). How you move is what you see: Action planning biases selection in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1755–1769. https://doi.org/10.1037/a0016798

Xu, X., Li, J., & Chen, H. (2022). Valence and arousal ratings for 11,310 simplified chinese words. *Behavior Research Methods*, *54*(1), 26–41. https://doi.org/10.3758/s13428-021-01607-4

Xu, Z., & Liu, D. (2024a). Body–object interaction effect in word recognition and its relationship with screen time in chinese children. *Reading and Writing*, *37*(4), 841–868. https://doi.org/10.1007/s11145-021-10238-2

Xu, Z., & Liu, D. (2024b). The role of body–object interaction in children's concept processing: Insights from two chinese communities. *Cognitive Processing*, *25*(3), 457–465. https://doi.org/10.1007/s10339-024-01185-1

Xue, J., Marmolejo-Ramos, F., & Pei, X. (2015). The linguistic context effects on the processing of body–object interaction words: An ERP study on second language learners. *Brain Research*, *1613*, 37–48. https://doi.org/10.1016/j.brainres.2015.03.050

Yap, M. J., Pexman, P. M., Wellsby, M., Hargreaves, I. S., & Huff, M. (2012). An abundance of riches: Cross-task comparisons of semantic richness effects in visual word recognition. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00072

Yap, M. J., Liow, S. J. R., Jalil, S. B., & Faizal, S. S. B. (2010). The malay lexicon project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, *42*(4), 992–1003. https://doi.org/10.3758/BRM.42.4.992

Yap, M. J., & Balota, D. A. (2015, September 1). Visual word recognition. In A. Pollatsek & R. Treiman (Eds.), *The oxford handbook of reading* (pp. 26–43). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199324576.013.4

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond coltheart's n: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

Yee, E., Chrysikou, E. G., Hoffman, E., & Thompson-Schill, S. L. (2013). Manual experience shapes object representations. *Psychological Science*, *24*(6), 909–919. https://doi.org/10.1177/0956797612464658

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review*, *23*(4), 1015–1027. https://doi.org/10.3758/s13423-015-0948-7

Yelland, G. W. (1994). Word recognition and lexical access. In *Encylopedia of language and linguistics* (Vol. 4). Pergamon Press.

Yu, A. B., Abrams, R. A., & Zacks, J. M. (2014). Limits on action priming by pictures of objects. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(5), 1861–1873. https://doi.org/10.1037/a0037397

Zeelenberg, R., & Pecher, D. (2016). The role of motor action in memory for objects and words. In *Psychology of learning and motivation* (pp. 161–193, Vol. 64). Elsevier. https://doi.org/10.1016/bs.plm.2015.09.005

Zhao, Y. C., Zhang, Y., Tang, J., & Song, S. (2021). Affordances for information practices: Theorizing engagement among people, technology, and sociocultural environments. *Journal of Documentation*, *77*(1), 229–250. https://doi.org/10.1108/JD-05-2020-0078

Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, *18*(5), 229–234. https://doi.org/10.1016/j.tics.2014.02.008

Zwaan, R. A. (2021). Two challenges to "embodied cognition" research and how to overcome them. *Journal of Cognition*, *4*(1), 14. https://doi.org/10.5334/joc.151

Appendices

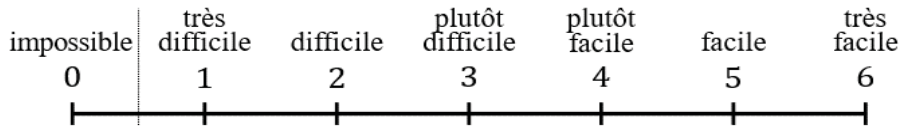
Appendix A
Supplementary materials for Chapter 3

Body-Object Interaction rating task instructions

Veuillez évaluer chaque mot selon la facilité avec laquelle le corps humain peut interagir physiquement et directement avec ce à quoi il se réfère.

Exemples: Il est facile d'interagir physiquement avec une CHAISE (s'asseoir dessus, la bouger, monter dessus), alors qu'il est plus difficile d'interagir avec un PLAFOND. Quant au SOLEIL, le corps ne peut pas interagir avec directement.

Vos évaluations devront être réalisées sur une échelle de 0 à 6, en choisissant la valeur que vous jugez pertinente. Une valeur de 0 indique que l'interaction physique avec le corps est impossible. Une valeur de 1 signifie qu'elle est très difficile et une valeur de 6 qu'elle est très facile. Les valeurs de 2 à 5 représentent des scores intermédiaires.





Il est important que vos évaluations ne soient basées que sur l'interaction physique, sans prendre en compte les différentes sensations (et leurs intensités) qui peuvent être impliquées. Par exemple, un objet qui est facile à voir ou à entendre mais avec lequel on ne peut pas interagir directement devrait recevoir un score faible.

Certains mots peuvent être ambigus. Il vous est demandé de les interpréter, quand c'est possible, comme des objets. Par exemple, le mot "orange" doit être interprété comme le fruit, et non la couleur.

Essayez d'être le plus précis possible dans vos évaluations, sans tout de même passer trop de temps sur chaque mot.

English translation

Please rate each word relative to the ease with which the human body can physically and directly interact with what it refers to.

Examples: It is easy to physically interact with a CHAIR (to sit on it, move it, stand on it), whereas it is more difficult to interact with a CEILING. As for the SUN, the body cannot interact with it directly.

Your ratings must be made on a scale from 0 to 6, by choosing the value that you judge pertinent. A value of 0 indicates that physical interaction with the body is impossible. A value of 1 means that it is very difficult and a value of 6 means that it is very easy. Values from 2 to 5 represent intermediate ratings.

[Scale: 0 - impossible, 1 - very difficult, 2 - difficult, 3 - somewhat difficult, 4 - somewhat easy, 5 - easy, 6 - very easy]

It is important that your ratings are only based on physical interaction, without considering the different sensations (and their intensities) which might be involved. For instance, an object which is easy to see or hear but with which it is not possible to directly interact should receive a low score.
Some words can be ambiguous. You should interpret them, when possible, as objects. For example, the word "orange" should be interpreted as the fruit, not the colour.

Try to be as precise as possible in your ratings, without spending too much time on each word.

Appendix B
Supplementary materials for Chapter 4

1 Semantic categories

The discussions in Chapter 4 and the analyses presented in the next section partly rely on the semantic categories to which rated items belong in the reviewed datasets. Category labels provided by eight studies were filtered and aggregated to obtain a general dataset (Banks & Connell, 2022; Brodeur et al., 2010, 2014; Miklashevsky, 2018; Moreno-Martínez et al., 2014; Navarrete et al., 2019; Ni et al., 2019; Stoinski et al., 2023), which was then used to classify the items in each reviewed rating study. Eleven broad categories were chosen to include objects that are typically associated with structural and/or functional manipulations (SM and FM respectively), as well as non-manipulable or potentially ambiguous items. Table B.1 presents the selected categories and their respective number of items in the final dataset.

Table B.1
Number of items in each semantic category

Category	*N*
Tool/utensil/supply	542
Instrument	118
Weapon	102
Clothing	264
Food	637
Body part	171
Animal	545
Vehicle	230
Furniture/appliance	183
Building	307
Profession	222
Total	3321

Note. 'utensil' refers to kitchen utensils, 'supply' to desk supplies, and 'appliance' to electrical appliances.

Studies reporting semantic categories differ in the labels that they have used and in the granularity of their categorisations. Tables B.2 through B.8 present the categories in each study that were grouped into those used in the present work. As Brodeur et al.'s (2014) study is an extension of Brodeur et al. (2010), the two datasets were combined in Table A.3.

Table B.2
Categories in Banks & Connell (2022) aggregated into those used in the current work

Category	Banks & Connell (2022)
Tool/utensil/supply	Carpenter's tool, gardening tool, tool, kitchen utensils
Instrument	Musical instrument, string instrument, wind instrument
Weapon	Weapon
Clothing	Clothing, hat
Food	Alcoholic drink, dairy product, nut, citrus fruit, fruit, green vegetable, vegetable, drug
Body part	Part of the body, part of the face
Animal	Animal, bird, bird of prey, breed of dog, farm animal, fish, four-legged animal, insect, rodent, snake, stinging insect, water bird
Vehicle	Boat, vehicle, two-wheeled vehicle, four-wheeled vehicle
Furniture/appliance	Furniture, living room furniture, kitchen appliance
Building	Building, human dwelling, religious building, room in house
Profession	Legal profession, healthcare profession, royal title

Table B.3
Categories in Brodeur et al. (2010, 2014) aggregated into those used in the current work

Category	Brodeur et al. (2010, 2014)
Tool/utensil/supply	Hand labour tool & accessory, kitchen & utensil, stationary & school supply
Instrument	Musical instrument
Weapon	War related weapon & item
Clothing	Clothing
Food	Food
Body part	Bodypart
Animal	Mammal, reptile, bird, insect, canine, sea mammal, feline, fish, crustacean
Vehicle	Vehicle
Furniture/appliance	Furniture
Building	Building infrastructure
Profession	/

Table B.4
Categories in Miklashevsky (2018) aggregated into those used in the current work

Category	Miklashevsky (2018)
Tool/utensil/supply	Tool power grip, tool precise grip
Instrument	/
Weapon	/
Clothing	Clothes
Food	Food
Body part	Body part
Animal	Animals, insects
Vehicle	Transport
Furniture/appliance	/
Building	Building
Profession	/

Table B.5
Categories in Moreno-Martínez et al. (2014) aggregated into those used in the current work

Category	Moreno-Martínez et al. (2014)
Tool/utensil/supply	Tools, kitchen utensils
Instrument	Musical instruments
Weapon	/
Clothing	Clothing
Food	Fruits, vegetables
Body part	Body parts
Animal	Animals, insects
Vehicle	Vehicles
Furniture/appliance	Furniture
Building	Buildings
Profession	/

Table B.6
Categories in Ni et al. (2019) aggregated into those used in the current work

Category	Ni et al. (2019)
Tool/utensil/supply	Office supplies, hand tools, kitchenware
Instrument	Musical instruments
Weapon	Weapons
Clothing	Clothes
Food	Fruits, nuts, vegetables
Body part	/
Animal	Aquatic, birds, fish, insects, mammals, reptiles
Vehicle	Vehicle
Furniture/appliance	Furniture
Building	/
Profession	/

Table B.7
Categories in Navarrete et al. (2019) aggregated into those used in the current work

Category	Navarrete et al. (2019)
Tool/utensil/supply	Desk material, kitchen utensils, tools
Instrument	Musical instruments
Weapon	Weapons
Clothing	Clothing
Food	Food, fruits, nuts, vegetables
Body part	Body parts
Animal	Animals, birds, insects, marine creatures
Vehicle	Vehicles
Furniture/appliance	Furniture
Building	Buildings
Profession	/

Table B.8
Categories in Stoinski et al. (2023) aggregated into those used in the current work

Category	Stoinski et al. (2023)
Tool/utensil/supply	Office supply, school supply, tool
Instrument	Musical instrument
Weapon	Weapon
Clothing	Clothing
Food	Food
Body part	Body part
Animal	Animal
Vehicle	Vehicle
Furniture/appliance	Kitchen appliance
Building	/
Profession	/

2 Manipulability ratings

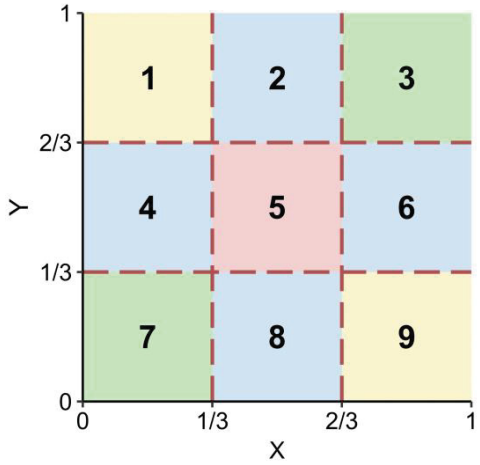
2.1 Introduction to the methods

The results discussed in Chapter 4 rely on three main types of graphical analyses[1]. The first one amounts to a simple distributional inspection (histogram) for a general assessment of the amount of midscale ratings. When available, we additionally plot the standard deviations (SD) as a function of the average ratings, as high SDs close to the middle of the scale can be indicative of disagreements as well. A second type of analysis aims to compare two variables (typically ratings from two studies using the same or similar instructions) through a scatter plot. The middle thirds of the scales are used to draw a grid, which allows a more guided reading of the relationship between two sets of ratings and to detect potential divergences. Figure B.1 illustrates the general approach that we use. If an item is in a green cell (3, 7), it means that it has been judged as either low or high in both cases. Similarly, if it has generated disagreement in both, it will fall in the red cell (5). The blue (2, 4, 6, 8) and yellow (1, 9) tiles are those containing conflictual observations. In the first case, it indicates that an item has been rated as low or high in one variable, but that it has been disagreed on in the other. The yellow cells represent a more extreme scenario in which participants have agreed to rate an item on opposite

[1] Note that our analyses are limited to studies providing English labels for their items to allow their interpretation and comparison.

sides of the scales. Note that the description provided here is only intended to convey the general idea behind the analysis and that the grid lines are not used as strict cut-offs. Each analysis requires a global assessment of the results and contextualised interpretations. For instance, if a large number of observations in a green cell are clustered at the edge of a grid line, this cannot be safely taken to indicate that the items were consistently rated on one end of the scales with high levels of agreement.

Figure B.1
Illustration of our bivariate ratings analysis

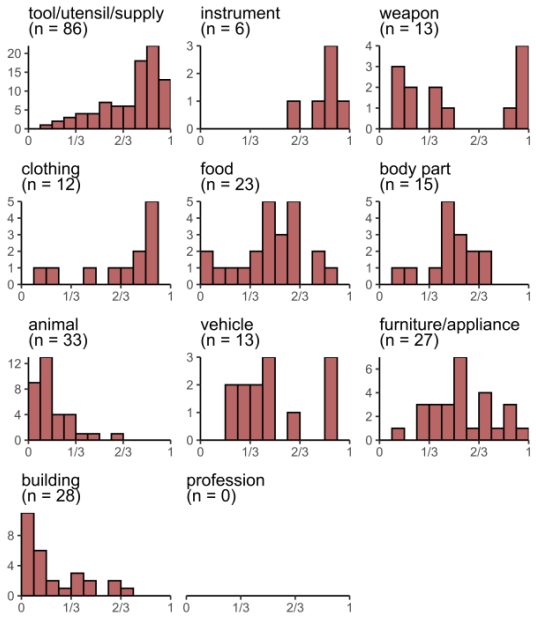


When two studies have used the same instructions to collect ratings, we should expect to find most of the common items in the green and red cells if their data are reliable. If two instructions offer valid assessments of a given dimension of manipulability, the ratings obtained through them should similarly be concentrated in these areas. Some noise will of course inevitably be present, resulting in some observations being scattered across the plot – especially considering the issue of measurement error (Chapter 3). The aim of this analysis is not to detect divergences for specific items, rather to assess the overall consistency of the data. This approach additionally serves as a screening method to identify potential issues with specific studies or dimensions.

The third analysis more directly aims to determine the types of objects that generate disagreement with a given instruction, and by extension to assess the latter's validity. For instance, if a scale proposed to capture FM does not yield high ratings for, e.g., tools and utensils, or if irrelevant objects also end up on its high end, then it cannot be reasonably considered a valid assessment of this dimension. This analysis is performed by first selecting

the items in each dataset that belong to the semantic categories presented in the previous section. When datasets provide SDs, we additionally remove the items with an average rating in the middle third of the scale that have a rescaled SD below 0.25. This ensures that only disagreed upon items are represented in this range of ratings. The distribution of ratings for each category is then plotted through histograms. An example is shown in Figure B.2 with the pantomime ratings provided by Guérard et al. (2015). We can see, for instance, that the majority of tools, utensils and supplies have been rated high, but that foods, among others, tend to be in the middle of the scale. Because the categories are not entirely homogeneous and might contain unusual classifications, we sometimes also manually inspect the ratings. Note that a major difficulty of this analysis is that some datasets contain only very few items in a given category, which might lead to misleading interpretations. We thus try to apply as much caution as possible when analysing and comparing the distributions.

Figure B.2
Example of category-wise rating distributions for Guérard et al.'s (2015) pantomime ratings



When no or only a single rating study is available for a given instruction, we additionally plot the summary statistics reported by experiments with a pilot rating phase for their different stimulus categories. Although significantly more limited, a combined assessment of their results can be informative as to the disagreements the instructions generate.

2.2 Structural manipulability

The pair-wise correlations between the identified SM ratings are presented in Table B.10. As a general observation, we can see that all variables are highly correlated with one another, except for Clarke and Lundington's (2018) ratings for the ease to grasp for moving (see Section 2.2.5). The discussions in this part are thus based on small differences between the instructions and rather bring out how different methodologies in rating studies affect the result.

Table B.10
Pair-wise Pearson correlations between structural manipulability ratings.

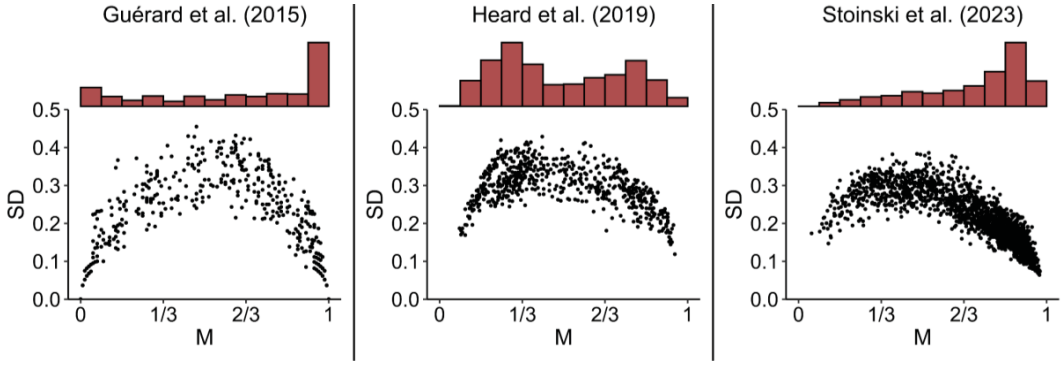
Study	EG (a)	EG (b)	EG (c)	LG (d)	LG (e)	EH (c)	EM (a)	EM (c)	EGM (f)
EG (a)		0.89	0.84	0.91	0.91	0.9	0.97	0.85	0.48
EG (b)	87		0.8	0.92	0.89	0.87	0.87	0.78	0.43
EG (c)	301	262		0.91	0.87	0.96	0.83	0.88	0.60
LG (d)	132	172	422		0.91	0.94	0.89	0.88	0.53
LG (e)	157	196	511	542		0.91	0.91	0.85	0.49
EH (c)	301	262	1854	422	511		0.89	0.92	0.56
EM (a)	559	87	301	132	157	301		0.86	0.46
EM (c)	301	262	1854	422	511	1854	301		0.54
EGM (f)	200	83	343	157	194	343	200	343	

Notes. The upper case letters correspond to rating instructions. EG: ease to grasp; LG: likelihood to grasp; EH: ease to hold; EM: ease to move; EGM: ease to grasp for moving. Lower case letters represent rating studies. a: Guérard et al. (2015); b: Heard et al. (2019); c: Stoinski et al. (2023); d: Amsel et al. (2012); e: Díez-Álamo et al. (2018); f: Clarke & Lundington (2018).
The lower triangle contains the number of common observations between two datasets. Significant Pearson correlation coefficients (α = .05) are reported in the upper triangle.

2.2.1 Ease to grasp

The distributional plots presented in Figure B.3 show that in both Guérard et al. (2015) and Stoinski et al. (2023), the ease to grasp ratings are heavily clustered on the high end of the scale. Heard et al.'s (2019) dataset contains relatively many more items towards the low end. However, most ratings are clustered around the edges of the middle of the scale. There are indeed notably little items on the extreme ends of the scale with SDs as low as in the other datasets. In contrast to the latter, this indicates that very few items were judged as consistently very easy or very difficult to grasp by all participants in Heard et al.'s (2019) study.

Figure B.3
Distributional histograms and the SDs against the means for the available ease to grasp ratings



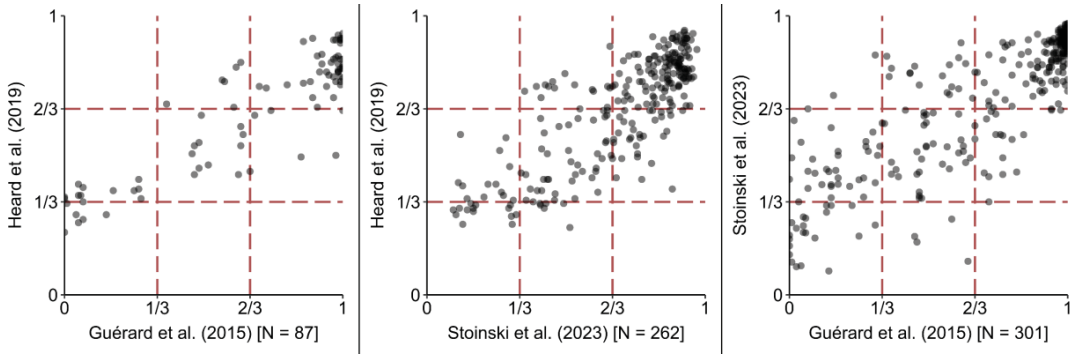
The left and middle scatter plots of Figure B.4 suggest that the items rated low in Guérard et al. (2015) and Stoinski et al. (2023) generated more disagreements in Heard et al. (2019). Similarly, some items rated high in Stoinski et al. (2023) appear to have been disagreed on in Heard et al. (2019). Note that Guérard et al. (2015) and Stoinski et al. (2023) normed object pictures, whereas Heard et al. (2019) included words only. This might have led to higher rates of disagreement due to the inherently more ambiguous nature of the stimuli. It is nevertheless unlikely that Heard et al.'s (2019) results are entirely due to this reason as other studies using words and similar instructions report relatively many more ratings on the extremes of the scale (e.g. Section 2.2.2).

The comparison of Guérard et al.'s (2015) and Stoinski et al.'s (2023) ratings also reveals a few divergences. Although a few items rated high in the latter study are found midscale in Guérard et al. (2015), a global assessment of the plot shows that there were overall more disagreements in Stoinski et al. (2023) for items that were agreed upon in Guérard et al.'s (2015) dataset – especially for those rated low. One possible explanation for this result is that Guérard et al. (2015) presented each participant with only one rating scale, while Stoinski et al. (2023) asked them to rate the same items on three very closely related dimensions: the ease to grasp, the ease to hold, and the ease to move. It is reasonable to suspect that this approach influenced participants' judgements – especially in the case of the ease to hold ratings. Indeed, the joint presence of these scales with different instructions suggests that they assess different characteristics. This might have encouraged participants to interpret them in unintended ways in order to distinguish them, thus leading to divergent judgements as a result. An interesting observation is that Stoinski et al.'s (2023) ease to hold ratings correlate more strongly with both

the ease to grasp ratings provided by the two other datasets, as well as with other SM dimensions than do their ease to grasp ratings (see Table B.10 and Section 2.2.3).

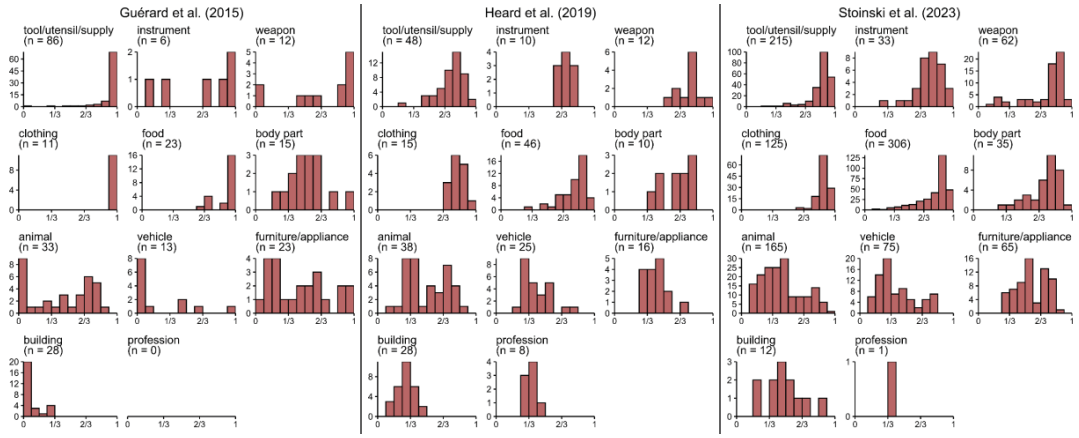
Figure B.4

Comparison of the ease to grasp ratings between datasets



Inspecting the rating distributions for the different categories (Figure B.5) reveals that most items expected to be graspable are indeed found on the high end of the scales in all studies (i.e. tools, utensils, supplies, instruments, weapons, clothes and foods). Although their number is small overall, some body parts can be found midscale, and mostly on the high end in Stoinski et al. (2023). Animals are distributed across the entire scale in all datasets. Those on the low end appear to be mainly big, sometimes dangerous animals, some of which participants are likely to have never or rarely encountered in person (e.g. *bear*, *hippopotamus*, *kangaroo*, *lion*, *rhinoceros*. Stoinski et al., 2023). Conversely, smaller and more common animals appear to be rated as easier to grasp (e.g. *butterfly*, *cat*, *fish*, *mouse*. Guérard et al., 2015). Despite this general pattern, where different animals are found on the scale is also quite variable across datasets. There thus appears to generally be uncertainty among participants as to how to rate them. This likely depends on how the instructions are interpreted individually. For instance, one participant might be judging whether they could in principle grasp an animal, while another also might consider the likelihood of encountering it or how difficult it would be to catch it. Furniture and appliances similarly appear to lead to considerable disagreements (e.g. *dishwasher*, *desk*, *microwave*, *shelf*. Stoinski et al., 2023). These results suggest that the instructions are ambiguous about whether the ratings should be based on the ease to grasp the objects in their entirety, or only parts of them.

Figure B.5
Category-wise rating distributions for the ease to grasp ratings

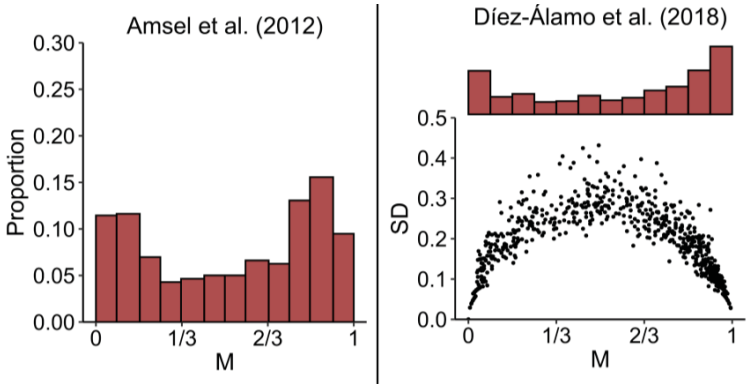


Interestingly, the datasets differ in their ratings for vehicles and buildings. There are relatively few vehicles in Guérard et al. (2015), but most are strongly clustered on the extreme low end of the scale. In contrast, almost none can be found to be this consistently rated low in both Heard et al. (2019) and Stoinski et al. (2023), and are more likely found towards the middle of the scale (e.g. *ambulance*, *bus*, *wheelchair*. Heard et al., 2019). The same observation can be made about buildings (e.g. *balcony*, *lighthouse*, *windmill*. Heard et al., 2019), although there are relatively few items in Heard et al. (2019) and Stoinski et al. (2023). These divergences are rather more difficult to explain, but the observed disagreements could be partly due to the ambiguity discussed above.

2.2.2 Likelihood to grasp

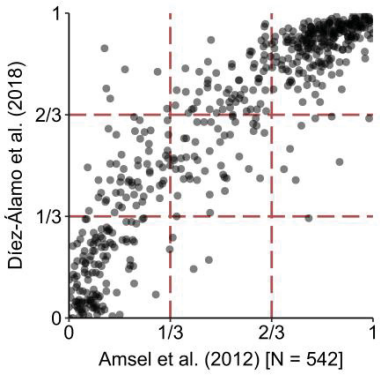
The ratings obtained with the likelihood to grasp instructions (Amsel et al., 2012; Díez-Álamo et al., 2018) appear to generally yield relatively few midscale items, with most of them being found towards the low and high ends of the scale (Figure B.6). The plots nevertheless suggest that the participants in Díez-Álamo et al.'s (2018) rating study tended to give more extreme responses, especially for items judged on the high end of the scale.

Figure B.6
Distributional histograms and the SDs against the means for the available likelihood to grasp ratings



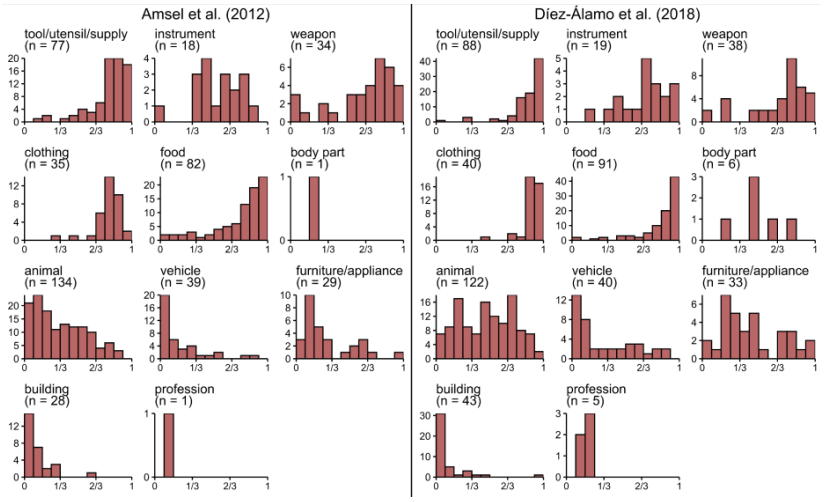
A direct comparison of the ratings (Figure B.7) clearly shows a non-linear relationship between the two datasets, suggesting that items generally received more responses on the high end of the scale in Díez-Álamo et al. (2018). We can notably see that some items in this dataset have midscale ratings, whereas they are rated low in Amsel et al. (2012). Similarly, several midscale items in the former dataset have high ratings in Díez-Álamo et al. (2018). It is difficult to determine the reason for this discrepancy as the two studies used the same instructions (although in different languages) and a highly overlapping set of words. The only major difference that we could identify is that Amsel et al. (2012) asked their participants to rate each item on 8 different dimensions, while Díez-Álamo et al. (2018) only presented them with a single rating scale "[w]ith the intention of minimizing potential effects of cross-dimension contamination" (p. 1634). It is difficult understand how this difference could have led to more 'high' ratings in the latter study given that the other dimensions used by Amsel et al. (2012) were not directly related to motor interaction. These results nevertheless strongly point to an influence of task settings on the ratings.

Figure B.7
Comparison of the likelihood to grasp ratings between datasets



The rating distributions for the different categories (Figure B.8) are similar to those observed for the ease to grasp ratings but reveal a few small differences. Instruments appear to have caused some disagreement, especially in Amsel et al. (2012. E.g. *banjo*, *flute*, *saxophone*). Compared to Heard et al.'s (2019) and Stoinski et al.'s (2023) ratings, vehicles and buildings also appear to be more consistently rated lower. As we have discussed, however, these two datasets might be slightly less reliable than Guérard et al.'s (2015) which display similarly low ratings for these categories. Once again, animals are distributed across the entire scale in both datasets. Furniture and appliances are also not clearly rated low or high, with several items found in the middle of the scale – especially in Díez-Álamo et al. (2018. E.g. *door*, *microwave*, *shelves*, *tv*).

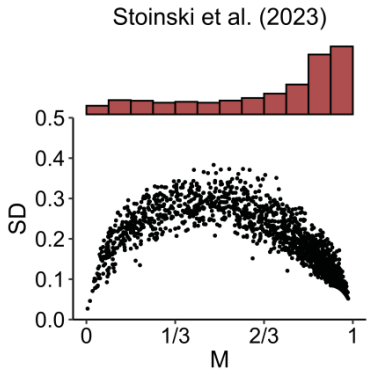
Figure B.8
Category-wise rating distributions for the likelihood to grasp ratings



2.2.3 Ease to hold

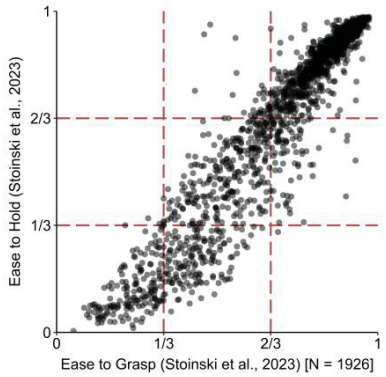
Stoinski et al.'s (2023) study is the only one reporting ease to hold ratings. Their distribution (Figure B.9) is very similar to that of the same study's ease to grasp ratings (Figure B.3). Slightly more items can nevertheless be seen in the low end of the scale, and those on the high end appear more strongly clustered at the scale's extremity. One experimental study (Rueschemeyer et al., 2010b) also reports summary statistics for two stimulus categories, namely functionally and volumentrically manipulable objects. These are discussed below along with the category-wise distributional analysis.

Figure B.9
Distributional histogram and the SDs against the means for the available ease to hold ratings



We have argued in Section 2.2.1 that Stoinski et al.'s (2023) ease to grasp ratings might have been influenced by concurrent judgements about the ease to hold. Comparing the two dimensions in the study reveals that a large portion of items found on the low end of the ease to hold scale generated disagreements when rated with the ease to grasp instructions (Figure B.10). Considering the stronger correlation of the ease to hold ratings with the other SM dimensions (Table B.10), these ratings thus appear to more reliably capture SM.

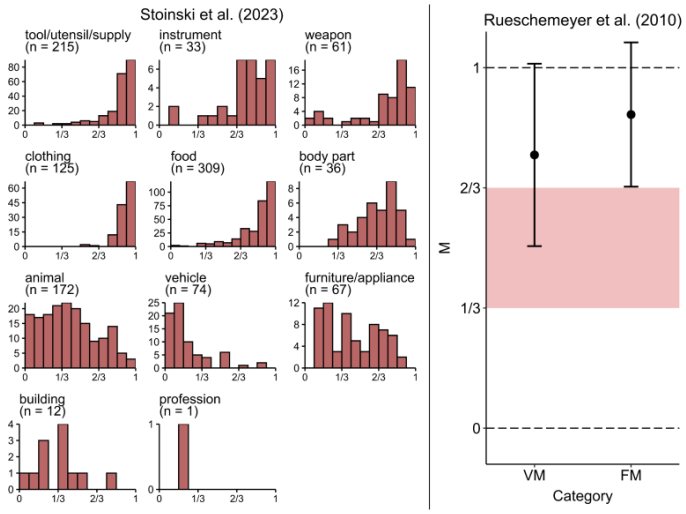
Figure B.10
Comparison of the ease to hold and ease to grasp ratings provided by Stoinski et al. (2023)



The different categories once again display very similar response distributions to the previous ratings (left panel of Figure B.11). Objects which can be expected to be SM are predominantly found on the high end of the scale (tools, utensils, supplies, instruments, weapons, clothes and foods). Animals, furniture and appliances also span the entire scale as in the previously described ratings. Nevertheless, note that more furniture and appliances appear on the low end of the scale compared to the ease to grasp ratings reported by the same study. Vehicles similarly appear to be rated more consistently on the low end. Body parts are generally found on the high end of the scale, with some generating disagreements (e.g. *eye*, *face*, *mouth*, *tongue*). The dataset contains only a few buildings, with some being in the middle of the scale (e.g. *fence*, *gate*, *tent*).

Figure B.11

Category-wise rating distributions for the ease to hold ratings (left) and the summary statistics of Rueschemeyer et al.'s (2010b) ease to hold ratings across two stimulus categories (VM: volumetrically manipulable, FM: functionally manipulable)



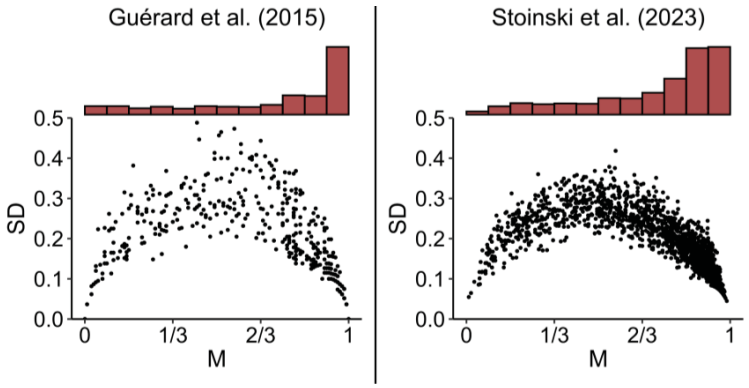
Note. The points correspond to each category's average ease to hold rating and the error bars represent one standard deviation around the means.

The right panel of Figure B.11 depicts the mean ease to hold ratings (and their standard deviations as error bars) for the two stimulus categories used in Rueschemeyer et al.'s (2010b) experiment. Volumetric manipulability is used synonymously with structural manipulability. Based on the study's methodology section, the ratings appear to have been collected to validate the experimental stimulus lists. In other words, the items were likely judged as SM and FM by the experimenters prior to the ratings study. The amount of variation that both categories display in their ratings is only possible if some items received ratings in the middle, and even on the low end of the scale. The stimuli used by this study are unfortunately not made available, thus limiting our conclusions. This observation nevertheless suggests that what was captured by the instructions did not entirely align with the experimenters' conception of SM – at least for a subset of the items.

2.2.4 Ease to move

The ease to move ratings were collected by Guérard et al. (2015) and Stoinski et al. (2023), whose other ratings were discussed above (ease to grasp and ease to hold). The rating distributions in the present case are highly similar to these other dimensions, with items strongly clustered on the high end of the scale (Figure B.12).

Figure B.12
Distributional histograms and the SDs against the means for the available ease to move ratings



A comparison of the ratings in the two datasets shows several divergent items (left plot of Figure B.13), but there does not appear to be a general pattern underlying them. For instance, items rated low in Stoinski et al. (2023) but high in Guérard et al. (2015) are *bench*, *cactus*, *keyhole*, *shelf*, *trampoline*. Those generating disagreements in Stoinski et al. (2023) but are rated high in Guérard et al. (2015) are similarly quite disparate (e.g. *bee*, *branch*, *faucet*, *light switch*, *mirror*, *sand*). The same goes for the other cells of the plot indicative of conflictual ratings, such as the midscale items in Guérard et al. (2015) that are low in Stoinski et al. (2023), e.g., *bed*, *cloud*, *couch*, *snowman*, *traffic light*. These differences thus seem largely random, which might be partly explained by the methodological differences discussed previously (see Sections 2.2.1 and 2.2.3).

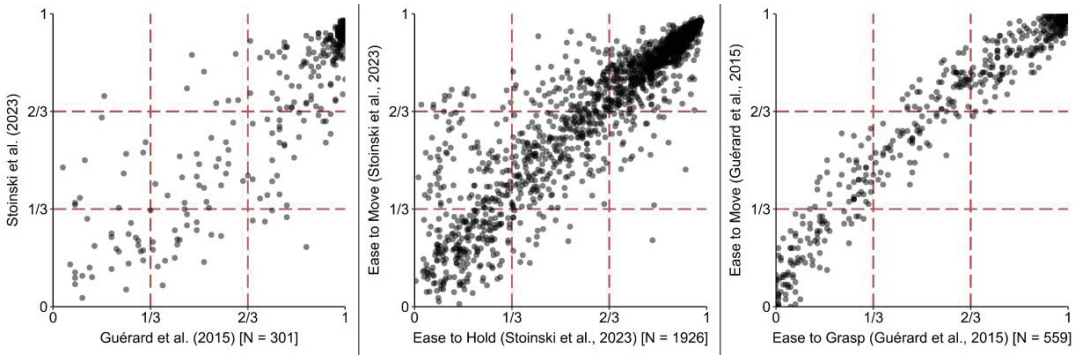
To further explore the ratings, the centre and right plots of Figure B.13 additionally plot the relationship of the ease to move ratings with the SM ratings provided by the same datasets. More specifically, the centre plot compares the ease to move ratings to those for the ease to hold in Stoinski et al. (2023), and the right plot shows their relationship to the ease to grasp ratings from Guérard et al. (2015). The two plots roughly display the same pattern of results, with the ease to move instructions resulting in higher (and mostly midscale) ratings for some items rated low for their ease to grasp or hold (case A), and some items that have generated disagreements on these dimensions being rated high for their ease to move (case B). We can also note a relatively more pronounced relationship between the two dimensions in Guérard et al. (2015).

In Stoinski et al. (2023), a manual inspection of the items in case A reveals that they are predominantly composed of vehicles (e.g. *ambulance*, *car*, *jetski*, *tractor*). In contrast, the same

case in Guérard et al. (2015) appears to contain generally large or fixed items not particularly related to a common category (e.g. *barbecue*, *cloud*, *couch*, *eye*, *electricity meter*, *statue (angel), tree trunk*). Case B in the former dataset appears to include mostly liquid or malleable items (e.g. *batter*, flour, *ink*, *lemonade*, *slime*, *soup*), and once again some vehicles (e.g. *cart*, *dirt bike*, *wheelchair*). Guérard et al.'s (2015) items in case B show a diverse set of items (e.g. *arm*, *burner*, *chair (office)*, *crucifix*, *keyhole*, *parachute*, *parrot*, *wheelchair*). These results generally suggest that objects that can move, either by themselves, using the hands, or through motorised means, can be rated high or generate disagreements on this dimension.

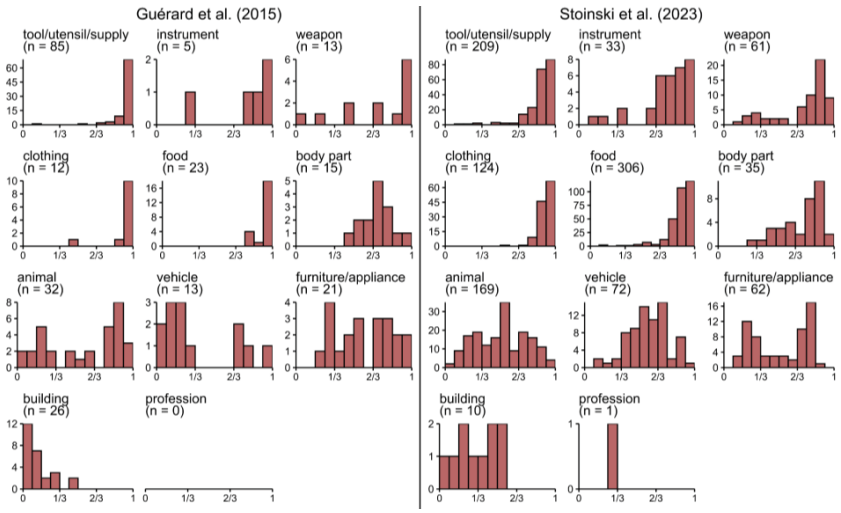
Figure B.13

Comparison of the ease to hold ratings between datasets (left), with Stoinski et al.'s (2023) ease to hold ratings (middle), and Guérard et al.'s (2015) ease to grasp ratings (right)



The category-wise rating distributions (Figure B.14) show that the items typically rated high in the previous scales continue to receive high ratings with the ease to move instructions (i.e. tools, utensils, supplies, instruments, weapons, clothes, foods). Body parts also display a similar pattern to the previous dimensions. Compared to the ease to grasp and the ease to hold dimensions in Guérard et al. (2015) and Stoinski et al. (2023) respectively, the ease to move ratings differ mainly for animals and vehicles which generally display higher values (including the middle of the scale). Considering that none of the two studies mention the use of hands to move the objects in their instructions, it is thus reasonable to suspect that some participants gave high ratings to objects that could move – irrespective of the means by which they would move.

Figure B.14
Category-wise rating distributions for the ease to move ratings



2.2.5 Ease to grasp for moving

The distribution of the ease to grasp for moving ratings reported by Clarke and Lundington (2018) shows that no items were consistently rated on the low end of the scale (Figure 15). The few items that can be found towards the low also have high SDs, suggesting that they did not have high agreement rates.

Figure B.15
Distributional histogram and the SDs against the means for the available ease to grasp for moving ratings

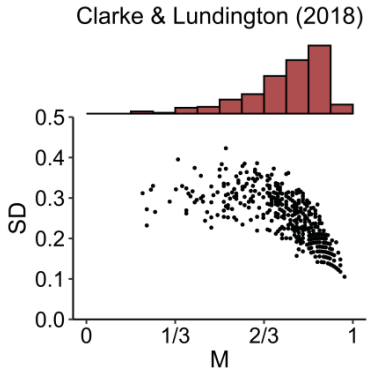
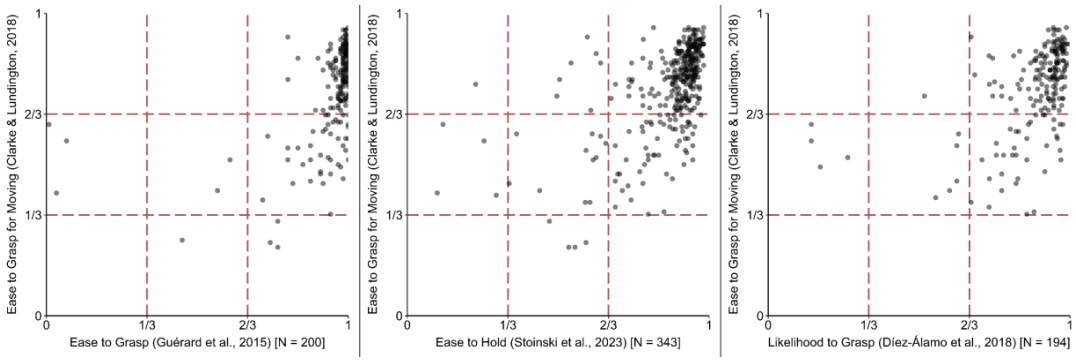


Figure B.16 shows a comparison of the current ratings with the most reliable ones obtained with the other SM instructions. The relationship of the ease to grasp for moving ratings

to the three other dimensions generally follows the same pattern: most items judged on the high end of the former scale have similarly high ratings for the ease to grasp (left plot. Guérard et al., 2015), the ease to hold (middle plot. Stoinski et al., 2023) and the likelihood to grasp (right plot. Díez-Álamo et al., 2018). However, several items judged high with the latter instructions appear to have been disagreed on with the ease to grasp for moving instructions. A manual inspection of these midscale items reveals that they contain a large number of items which are unambiguously easy to grasp and move (e.g. *axe*, *cigarette*, *feather*, *light bulb*, *pine cone*, *pliers*, *saw*, *vase*, *whisk*).

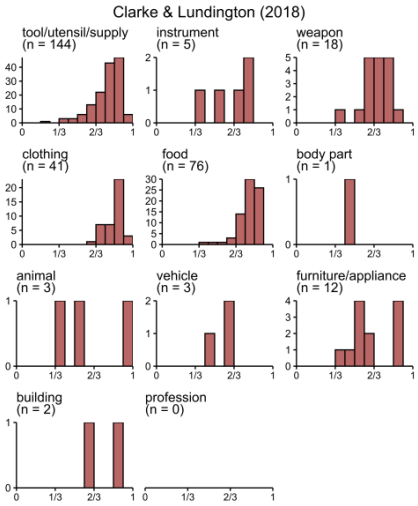
Figure B.16

Comparison of the ease to grasp for moving ratings with the ease to grasp ratings from Guérard et al. (2015. Left), the ease to hold ratings from Stoinski et al. (2023. Middle) and the likelihood to grasp ratings from Díez-Álamo et al. (2018. Right)



Given that most items in the dataset are rated on the high end of the scale, a closer look at the categories-wise distributions unsurprisingly shows that most have received high ratings (Figure B.17). Once again, however, a manual inspection reveals that several tools, utensils and supplies display disagreements (e.g. *bottle*, *chisel*, *clamp*, *cutting board*, *drill*, *match*, *scissors*). These instructions point to an issue with the validity of the current ratings. In light of the data presented in the previous sections, however, it is highly unlikely that the simple addition of the action's goal (i.e. moving) to the instructions could have resulted in such disagreements. An important point regarding Clarke and Lundington's (2018) study is that the rated item set was almost exclusively composed of SM objects. Additionally, participants were explicitly encouraged to "utilize the full range of the scale when rating objects" (p. 613). In the absence of objects that are not clearly difficult to grasp for moving, it is thus possible that some participants calibrated their ratings relative to the items they were presented.

Figure B.17
Category-wise rating distributions for the ease to grasp for moving ratings



2.3 Functional manipulability

As discussed in the main text, FM has been operationalised through a much more diverse set of instructions than SM. Table B.11 presents the pair-wise Pearson correlations between the FM ratings reported by different rating studies. We can already observe that the relationships between the different dimensions are much less pronounced compared to SM ratings, indicating that the instructions vary to some extent in what they assess.

Table B.11

Pair-wise Pearson correlations between functional manipulability ratings

Study	AA(a)	HAA(b)	HAA(c)	HAA(d)	M(e)	HN(f)	HN(g)	HN(h)	HN(i)	HN(j)	P(k)	P(l)	P(m)	Pm(n)	Pm(o)	Pm(p)	GUD(q)	GUD(r)	NoA(r)	NoA(n)	NoA(p)
AA(a)		0.51	0.65	0.49	0.68	0.69	0.67	0.78	0.79	0.73	0.84	0.66	0.78	0.81	ns	0.48	0.53	ns	ns	ns	ns
HAA(b)	50		0.71	0.7	0.75	0.78	0.83	0.62	0.76	0.9	0.67	0.52	0.75	0.77	ns	0.47	0.53	ns	ns	0.36	ns
HAA(c)	154	506		0.79	0.53	0.71	0.67	0.62	0.69	0.75	0.6	0.31	0.57	0.62	0.16	0.44	0.29	-0.23	0.2	0.35	0.29
HAA(d)	55	137	909		0.5	0.6	0.66	0.68	0.78	0.78	0.5	ns	0.43	0.56	0.24	0.35	ns	-0.42	-0.35	0.2	0.32
M(e)	111	380	1367	339		0.86	0.8	0.64	0.73	0.82	0.74	0.48	0.68	0.71	ns	0.55	0.49	ns	ns	0.37	0.15
HN(f)	26	24	107	37	48		0.98	0.82	0.91	0.98	0.7	0.85	0.6	0.82	ns	ns	ns	ns	ns	ns	ns
HN(g)	61	87	326	101	185	103		0.85	0.93	0.96	0.7	0.4	0.68	0.81	0.33	ns	0.34	ns	0.49	0.21	ns
HN(h)	78	109	592	135	260	100	199		0.85	0.86	0.71	0.45	0.62	0.81	0.18	0.24	0.27	-0.33	ns	ns	ns
HN(i)	51	86	437	127	191	24	97	144		0.91	0.82	0.41	0.63	0.82	ns	ns	0.33	-0.4	0.36	ns	ns
HN(j)	55	70	299	86	160	108	327	191	87		0.68	0.44	0.7	0.83	0.25	0.32	0.35	ns	ns	ns	ns
P(k)	77	79	256	101	178	38	123	151	105	105		0.67	0.67	0.83	0.27	0.53	0.38	-0.32	0.27	0.2	ns
P(l)	47	45	373	89	168	26	108	140	89	95	130		0.48	0.73	0.36	0.7	0.33	-0.18	0.17	0.26	-0.26
P(m)	104	143	641	175	320	41	156	216	154	139	217	143		0.76	0.3	0.64	0.49	-0.38	ns	0.53	ns
Pm(n)	55	64	363	97	155	27	91	118	82	83	133	200	337		0.55	0.49	0.32	-0.21	0.43	0.59	ns
Pm(o)	47	45	373	89	168	26	108	140	89	95	130	606	143	200		0.37	ns	ns	ns	0.2	ns
Pm(p)	56	91	621	94	228	27	73	123	71	70	77	83	129	87	83		ns	ns	ns	0.37	0.47
GUD(q)	41	31	164	53	98	21	67	84	50	58	113	131	92	87	131	43		0.46	0.52	0.34	ns
GUD(r)	16	17	162	31	59	5	33	40	37	27	51	212	34	115	212	31	58		0.45	ns	ns
NoA(r)	16	18	175	36	66	5	35	42	41	29	55	228	39	122	228	35	61	278		0.58	ns
NoA(n)	55	64	363	97	155	27	91	118	82	83	133	200	337	560	200	87	87	115	122		0.53
NoA(p)	56	91	621	94	228	27	73	123	71	70	77	83	129	87	83	621	43	31	35	87	

Notes. The upper case letters correspond to rating instructions. AA: action association; HAA: hand/arm action association; M: manipulability; HN: hand necessity for function; P: pantomime (original); Pm: pantomime (modified); GUD: grasp-use dissimilarity; NoA: number of actions.

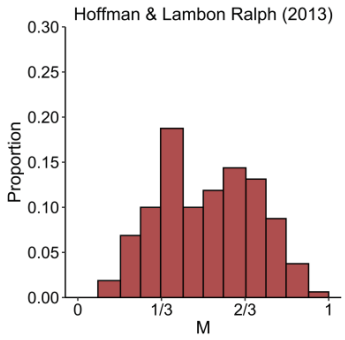
Lower case letters represent rating studies. a: Hoffman & Lambon Ralph (2013); b: Binder et al. (2016); c: Lynott et al. (2020); d: Repetto et al. (2023); e: Medler et al. (2005); f: Moreno-Martínez et al. (2011); g: Moreno-Martínez & Montoro (2012); h: Moreno-Martínez et al. (2014); i: Miklashevsky (2018); j: Navarrete et al. (2019); k: Magnié et al. (2003); l: Brodeur et al. (2010); m: Brodeur et al. (2014); n: Guérard et al. (2015); o: Clarke & Lundington (2018); p: Heard et al. (2019); q: Salmon et al. (2010); r: Lagacé et al. (2013).

The lower triangle contains the number of common observations between two datasets. Significant Pearson correlation coefficients ($\alpha = .05$) are reported in the upper triangle.

2.3.1 Action association

We can see in Figure B.18 that the action association ratings collected by Hoffman and Lambon Ralph (2013) are heavily clustered around the middle of the scale. Indeed, 55% of all items can be found in its middle third, whereas almost none were consistently judged as very high or very low.

Figure B.18
Distributional histogram for the available action association ratings



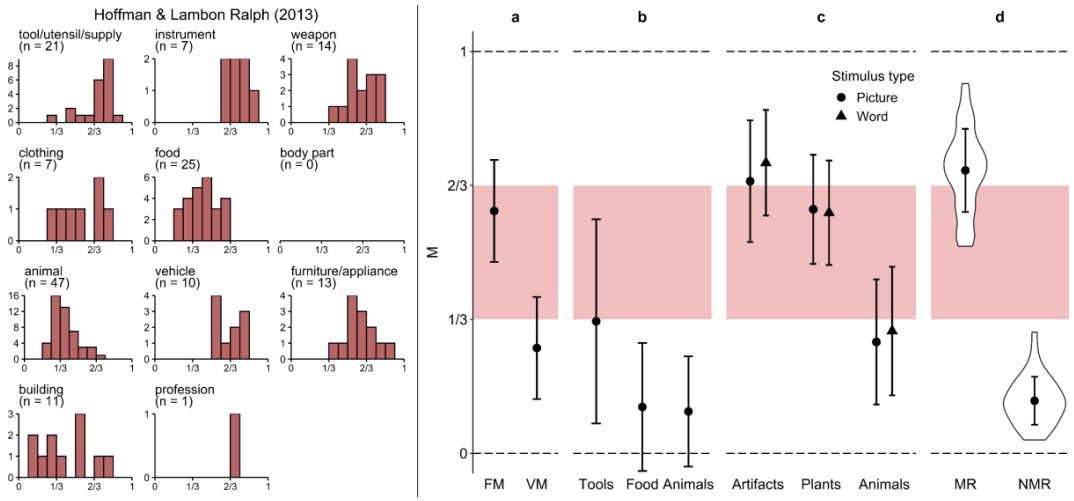
The left panel of Figure B.19 reveals that objects that would typically be considered FM, (tools, utensils, supplies, instruments, weapons) are towards the high end of the scale. However, some are disagreed on (e.g. *brush*, *clarinet*, *gun*, *machete*, *spear*, *tuba*) and most are close to the middle of the scale. The other categories display a similar pattern with part of the items in the middle of the scale, and the rest found mostly on one of the ends. The disagreements observed for animals (e.g. *duck*, *elephant*, *moth*) and vehicles (e.g. *helicopter*, *subway*, *truck*) notably suggest that the instructions are too generic to be considered a valid assessment of FM as they can be interpreted as referring to any type of action. In such a rating context, most objects could be potentially thought of as being somewhat associated with action (e.g. animals and vehicles can move, a rock can be thrown, we perform actions in front of mirrors, a mountain can be climbed). This leaves excessive room for interpretation and thus likely leads to disparate judgements and to mostly midscale ratings.

A few experiments have collected action association ratings in pilot studies and report summary statistics for their stimulus categories (Carota et al., 2012; Gainotti et al., 2013; Proverbio et al., 2011; Rueschemeyer et al., 2010b). Magri et al. (2021) have provided more detailed item-level ratings – although not the names of the items. The data from each study are plotted in the right panel of Figure B.19. Proverbio et al.'s (2011) were omitted as our reading of the summary statistics might be misleading for their ratings that were collected on a shorter,

3-point scale. Similarly to Hoffman and Lambon Ralph's (2013) ratings, none of the categories studied by these experiments appear to contain items rated very high on the scale. Their comparison additionally shows some important inconsistencies. For instance, FM items in Rueschemeyer et al. (2010b. Plot a) appear to mostly have midscale ratings and to be slightly distributed towards the high end of the scale. In Carota et al. (2012. Plot b), however, tools surprisingly have an average rating on the lower end of the middle of the scale. Their standard deviation additionally suggests that most items were rated low and that a few were rated high. The two other studies (plot c: Gainotti et al., 2013; plot d: Magri et al., 2021) are a little more difficult to interpret as their categories are relatively more general. We can nevertheless note that those that would be expected to contain FM objects (artifacts and motor-related, respectively) have distributions close to the middle of the scale with no items on its extreme high end. Other categories also differ across studies. Animals and foods were rated low in Carota et al. (2012), but can be mostly found close or in the middle of the scale in Gainotti et al. (2013)[2].

Figure B.19

Category-wise rating distributions for Hoffman and Lambon Ralph's (2013) action association ratings (left) and their summary statistics for different stimulus categories (right) in (a) Rueschemeyer et al. (2010), (b) Carota et al. (2012), (c) Gainotti et al. (2013), and (d) Magri et al. (2021)



Notes. FM: functionally manipulable, VM: volumetrically manipulable, MR: motor-related, NMR: non-motor-related.
Points represent the average rating for each category and the error bars display one standard deviation around the mean.

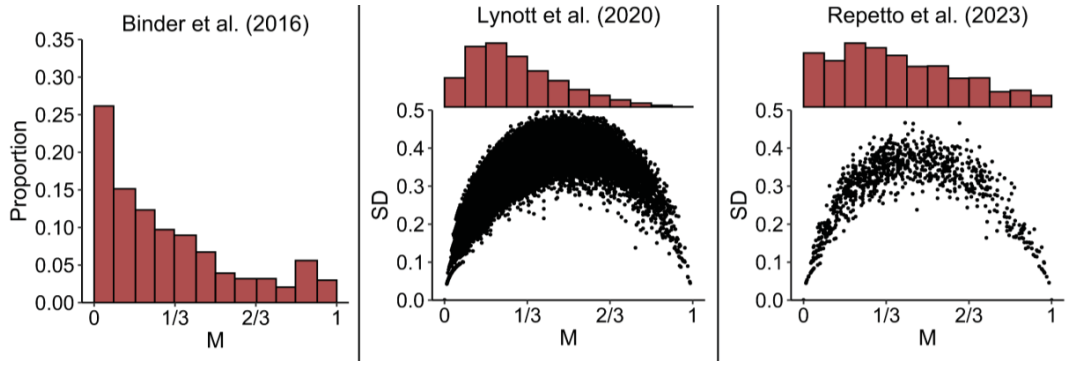
[2] The 'plants' category in Gainotti et al. (2013) was composed of flowers, fruits and vegetables.

To our knowledge, two of the experiments with a pilot rating phase used this data for stimulus validation – and not stimulus selection (Carota et al., 2012; Rueschemeyer et al., 2010b). As the items were pre-selected, one would expect their categories containing FM objects (tools and functionally manipulable objects, respectively) to receive high ratings on the scale. However, both contain significant disagreements – and even low ratings – which shows that the instructions did not align with the criteria used by the researchers. Overall, these observations thus strongly suggest that the action association instructions are not a valid assessment of FM and that they result in substantially noisy ratings.

2.3.2 Hand/arm action association

The hand/arm action association ratings generally display positively skewed distributions (Figure B.20) indicating that relatively few items were consistently rated on the high end of the scale.

Figure B.20
Distributional histograms and the SDs against the means for the ease to grasp ratings



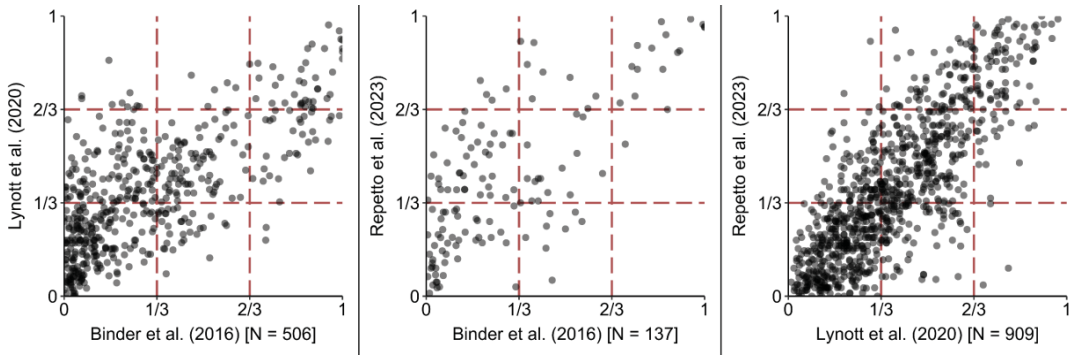
A comparison of the ratings reported by the three datasets reveals some divergences between Binder et al.'s (2016) data and those from Lynott et al. (2020) and Repetto et al. (2023). As can be observed on the left and middle plots of Figure B.21, several midscale items in Lynott et al. (2020) and Repetto et al. (2023) are found on the low end of the scale in Binder et al. (2016). In the case of Lynott et al. (2020) we can also see that several midscale items are rated high in Binder et al. (2016). Some low items in the former study are also disagreed on in Binder et al. (2016). In contrast, the comparison of Lynott et al.'s (2020) and Repetto et al.'s (2023) ratings shows a much more linear – although noisy – relationship between the ratings.

There are some small differences between the instructions used by Binder et al. (2016) and those of Lynott et al. (2020) and Repetto et al. (2023). The former authors notably asked

about the extent to which participants think of the items as "being associated with […] actions using the arm, hand, or fingers" (in online supplementary materials), whereas the two other studies asked about the extent to which participants experience each item by performing hand/arm actions. Variable degrees of experience with the objects among participants could have led to disparate ratings and thus to midscale items. On the other hand, Binder et al.'s (2016) instructions inquire more generally about the association to manual actions irrespective of direct experience, and might have thus resulted in lower disagreements. Another potential confound is the inclusion of action verbs in Lynott et al.'s (2020) and Repetto et al.'s (2023) studies, which could have biased the judgements for nouns (see below).

Figure B.21

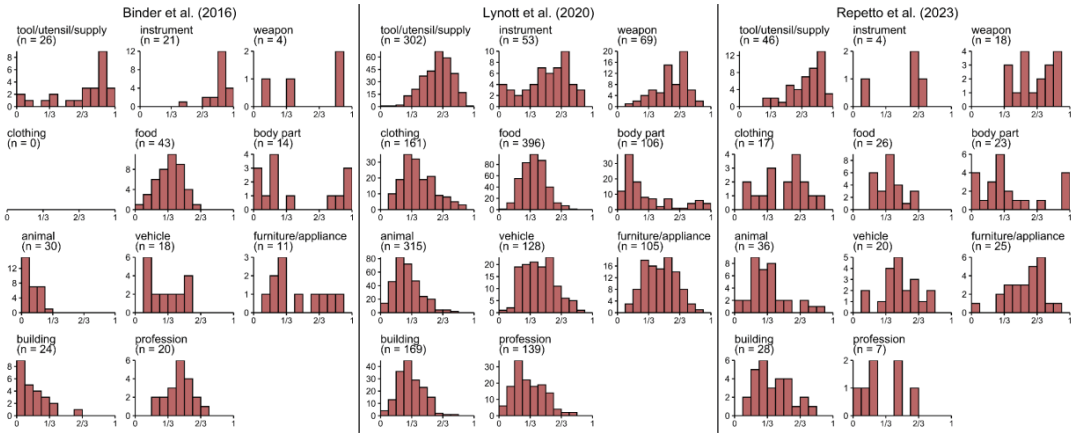
Comparison of the hand/arm action association ratings between datasets



The category-wise rating distributions mostly confirm the differences discussed above (Figure B.22). Objects that can be expected to be FM (tools, utensils, supplies, instruments) are clustered on the high end of the scale in Binder et al. (2016). Conversely, non-FM items such as animals and buildings are on the low end of the scale. However, the results of Lynott et al. (2020) show that almost all categories contain a significant number of midscale items, even for FM objects (e.g. tools, utensils, supplies, instruments, animals, vehicles, buildings). Repetto et al.'s (2023) ratings mostly display similar distributions to those of Lynott et al. (2020). Tools, utensils and supplies are nevertheless found relatively higher on the scale. Finally, foods and profession names appear to equally generate disagreements in all datasets.

Figure B.22

Category-wise rating distributions for the hand/arm action association ratings



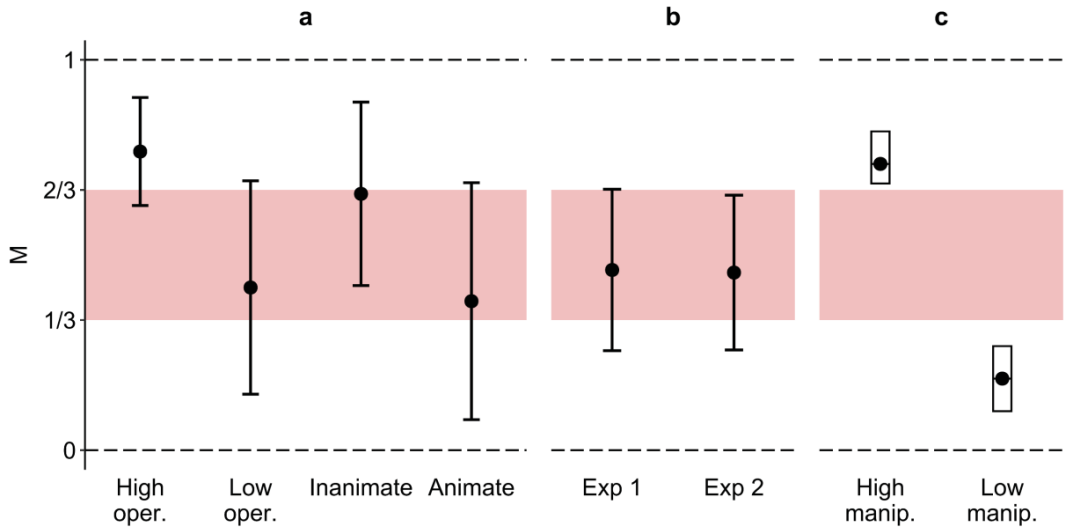
The analyses presented above suggest that Lynott et al.'s (2020) and Repetto et al.'s (2023) ratings are relatively less reliable than those reported by Binder et al. (2016). On the one hand, the two former studies focused on participants' experience with the rated items which can result in considerable disagreements due to individual differences. Such instructions can be highly pertinent in experimental contexts in which the same participants performing a task also provide ratings for their experience of hand/arm actions with the objects. However, their use for norming FM appears rather limited as they appear to introduce considerable noise to the ratings. On the other hand, it is unlikely that the disagreements observed in these two studies – particularly in Lynott et al. (2020) – are entirely due to individual differences and to the instructions. For instance, several very common FM objects are surprisingly found midscale in this study (e.g. *crayon*, *highlighter*, *kettle*, *mug, pencil case*, *razor*, *screw*, *toothbrush*). As mentioned earlier, some of the disagreements might have been caused by the presence of action words. Indeed, these appear to have received very high ratings on the scale (e.g. *pointing*, *touching*, *writing*, *carry*, *applaud*) and it is reasonable to suspect that they served as strong anchors on the high end of the scale. They likely influenced different participants at a different extent, depending partly on their interpretation of the instructions and the order in which items were presented.

2.3.3 Manipulability

As we have discussed in the main text, the instructions grouped under this category present some important differences in their wording. Regarding the studies that will be discussed below, Howard et al. (1995) and Desai et al. (2016) have used similar instructions, generally asking participants to rate how manipulable objects are. The summary statistics for the ratings collected in Haddad et al.'s (2023) and Kellenbach et al.'s (2003) pilot studies will not be presented as they might be misleading in the context of our current method. In the former case, the authors only report median values for their stimulus categories, which is not informative about the overall distribution of their ratings. On the other hand, Kellenbach et al. (2003) used a 3-point scale for their ratings, which likely does not lend itself to the same reading as the scales discussed here. The only available rating study for this category of instructions was presented by Medler et al. (2005) who further defined "manipulation as a physical action done to an object by a person" (online materials). Finally, Pecher et al. (2013) used a different definition of manipulability, presented as the "extent to which actions are carried out with an object" (our translation of the original ratings in Dutch, personal communication, January 15, 2024).

Figure B.23 presents the summary statistics from (a) Howard et al.'s (1995), (b) Desai et al.'s (2016) and (c) Pecher et al.'s (2013) pilot rating studies. The four categories in (a) correspond to two different groupings of the same stimuli as either high and low in operativity or as animate and inanimate. We can see that high operativity items (which supposes their manipulability) are towards the high end of the scale but likely included some midscale items. The other categories, however, appear to contain a large number of midscale items. The large standard deviations additionally suggest that low operativity, animate and inanimate items were distributed across the entire scale. Unfortunately, Desai et al. (2016) have only provided summary statistics for the whole stimulus lists of their two experiments. It is thus not possible to determine how different categories are distributed. The data nevertheless shows that the majority of the items must have been in the middle of the scale. Finally, Pecher et al.'s (2013) study reports a more convenient description with the range of ratings for each category. We can see that high and low manipulability objects are well separated, respectively on the high and low ends of the scale. We can nevertheless note that none of the high manipulability objects are on the extreme end, i.e. no item was consistently rated on the high end of the scale by all participants.

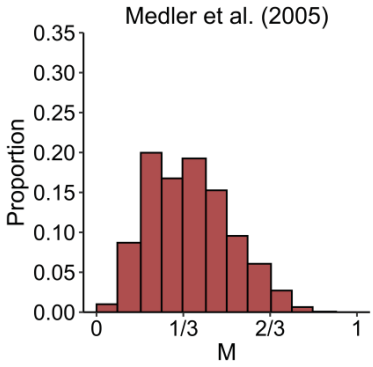
Figure B.23
Summary statistics of manipulability ratings for different stimulus categories in (a) Howard et al. (1995), (b) Desai et al. (2016), and (c) Pecher et al. (2013)



Notes. oper.: operativity, manip.: manipulability.
Points represent the average rating for each category and the error bars display one standard deviation around the mean. The boxes in (c) show the range of the ratings

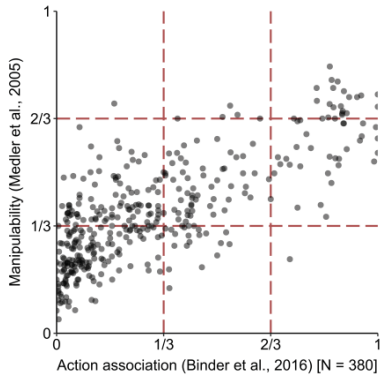
The ratings reported by Medler et al. (2005) are presented in Figure B.24. Similar to the action association ratings (Section 2.3.1), we can see that a large portion of the items are in the middle of the scale (50%). There are additionally no items consistently rated on its high end, and very few on the low side. This observation once again suggests that the instructions used by these authors led to significant disagreements among participants.

Figure B.24
Distributional histogram for the available manipulability ratings



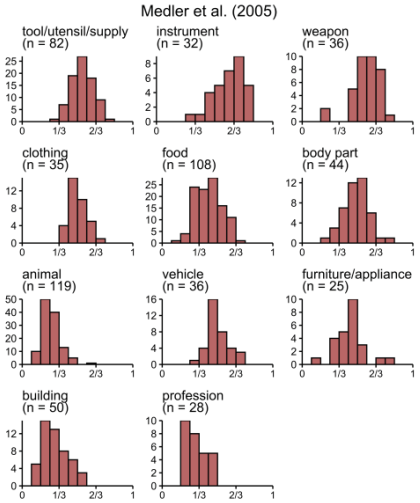
Binder et al.'s (2016) hand/arm action association ratings are arguably the most reliable among the FM variables reviewed so far. Their comparison with Medler et al.'s (2005) manipulability ratings (Figure B.25) shows that several items rated both high and low for their association to hand/arm actions generated disagreements in Medler et al. (2005).

Figure B.25
Comparison of Medler et al.'s (2005) manipulability ratings with the hand/arm action association ratings from Binder et al. (2016)



Finally, Figure B.26 confirms that the items from virtually all analysed categories generated disagreements. Only animals, buildings and professions appear to have received relatively more responses on the low end of the scale. Items falling in this category nevertheless display ratings close to the middle of the scale as well and likely generated disagreements as well. The manipulability ratings provided by Medler et al. (2005) are thus not informative for FM as they do not appear to be reliably distinguishing these items. Overall, the manipulability instructions appear to cause significant disagreements when manipulability is not or inappropriately defined. Pecher et al.'s (2013) data suggest that their definition captures FM somewhat more reliably. Unfortunately, the lack of available item-level data and the observation that most items classified as highly manipulable remain close to the middle of the scale call for a cautionary stance towards their instructions.

Figure B.26
Category-wise rating distributions for the manipulability ratings

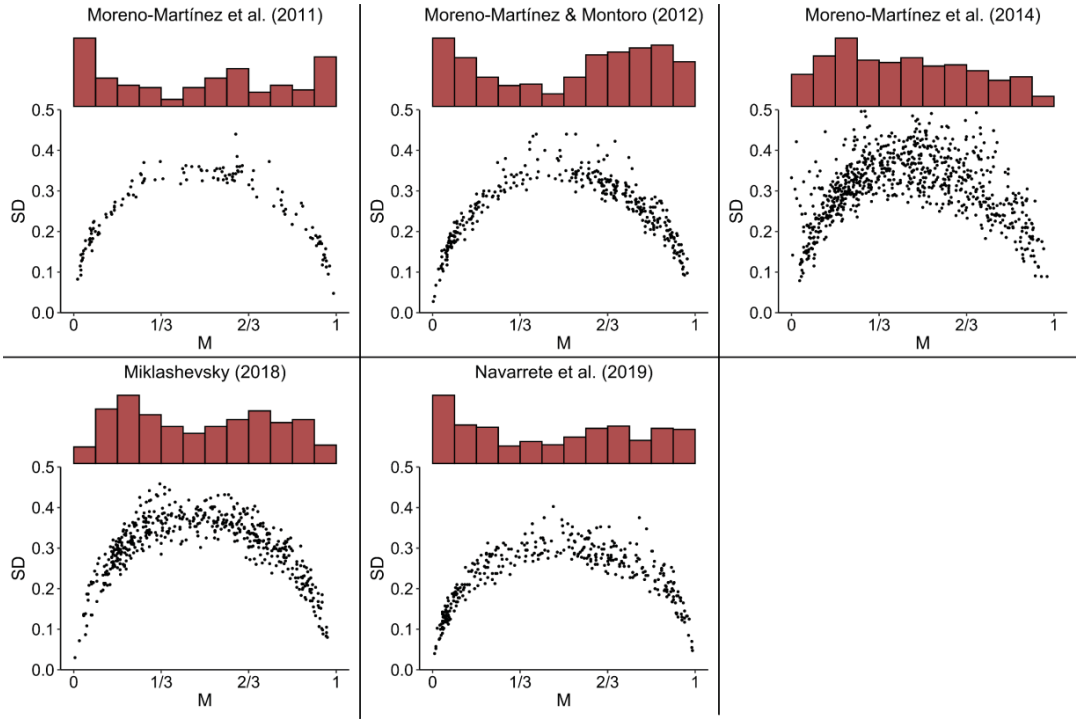


2.3.4 Hand necessity for function

To our knowledge, five rating studies have collected hand necessity for function ratings. An inspection of their distributions and SDs reveals that the data reported by Moreno-Martínez et al. (2014) display some inconsistencies (Figure B.27). Several items indeed have average ratings outside of the scale's range as reported in the study (e.g. *scythe* has an average rating of 0.76, on a [1, 5] scale) and SDs that would be impossible to observe on an ordinal scale (e.g. *stone mason chisel*, $M = 1.074$, $SD = 1.69$). These observations point to an error during the computation of the summary statistics. In the following analyses, we will thus exclude this study, as well as the data from Moreno-Martínez et al. (2011) as the number of normed items is relatively low ($N = 140$). Despite some apparent midscale items in all datasets, Figure B.27 shows that the hand necessity for function instructions roughly result in bimodally distributed ratings. This means that they are able to capture items on both the low and the high ends of the scale.

Figure B.27

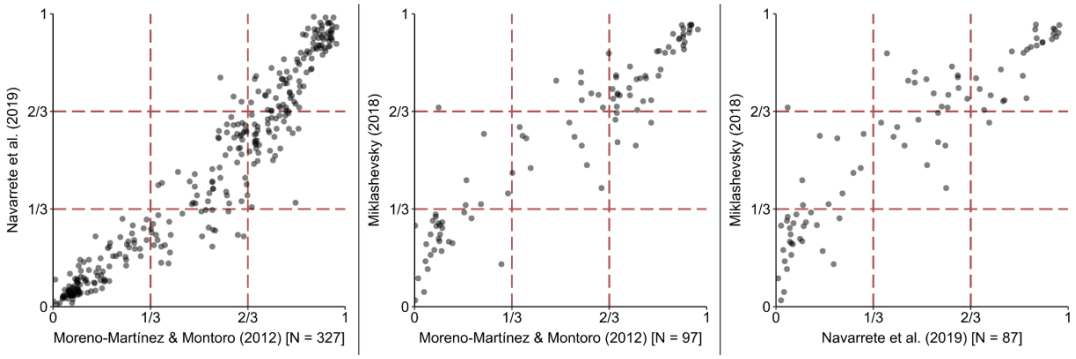
Distributional histograms and the SDs against the means for the available hand necessity for function ratings



Comparing the data provided by the different datasets shows that there are no major differences between them (Figure B.28). Given the low number of items in common between Miklashevsky (2018) and Moreno-Martínez and Montoro (2012), and between the former study and Navarrete et al. (2019), it is likely the few divergent ratings are a result of measurement error. We can nevertheless note that items rated as low in both Moreno-Martínez and Montoro (2012) and Navarrete et al. (2019) appear to have caused relatively more disagreements in Miklashevsky (2018).

Figure B.28

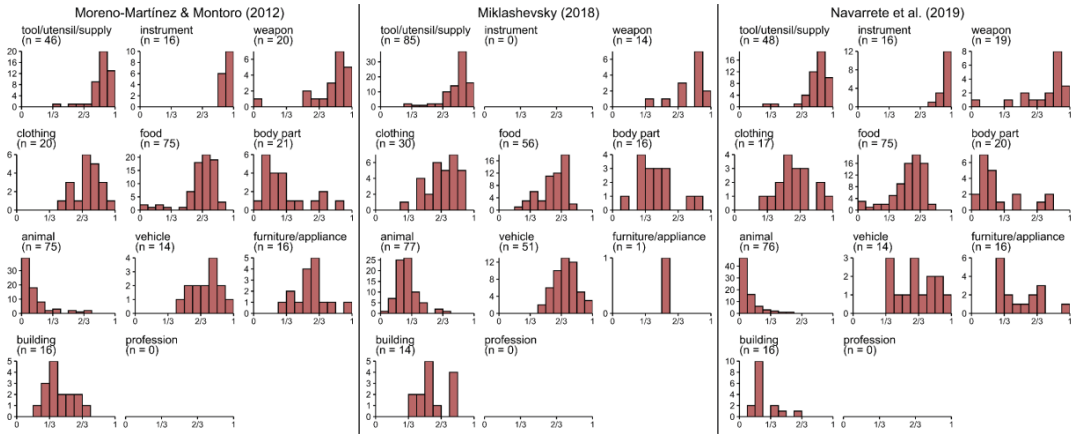
Comparison of the hand necessity for function ratings between datasets



The category-wise rating distributions (Figure B.29) reveal that the hand necessity for function instructions reliably capture typically FM items (i.e. tools, utensils, supplies, instruments, weapons) on the high end of the scale. Conversely, animals and body parts appear largely on the low end of the scale. Despite ratings leaning towards the high end of the scale, we can also see that several clothes, foods and vehicles (e.g. *avocado*, *bus*, *coat*, *shirt*, *tomato*, *train*. Navarrete et al., 2019) have led to disagreements among raters. Few buildings, furniture items and appliances are present in these datasets but also seem to cause some amount of disagreement (e.g. *bed*, *church*, *couch*, *factory*, *table*, *tower*. Moreno-Martínez & Montoro, 2012).

Figure B.29

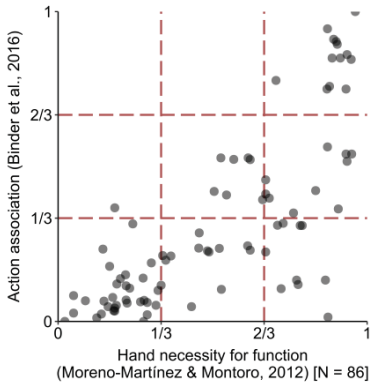
Category-wise rating distributions for the hand necessity for function ratings



Finally, we can compare the hand necessity for function ratings (Moreno-Martínez & Montoro, 2012) with those for hand/arm action association reported by Binder et al. (2016). Despite a low number of items in common, Figure B.30 shows that several midscale and low

items in Binder et al. (2016) are rated high for their hand necessity for function in Moreno-Martínez & Montoro (2012). Some items rated low in the former case can also be found in the middle of the scale in Martínez & Montoro (2012). In all of these cases, the majority of items appear to refer to foods, furniture and vehicles (e.g. *bed*, *bus*, *car*, *carrot*, *cucumber*, *chair*, *scooter*, *table*). Overall, the hand necessity for function instructions thus reliably capture typically FM objects, but appear to be ambiguous for several other types of objects.

Figure B.30
Comparison of Moreno-Martínez and Montoro's (2012) hand necessity for function ratings with the hand/arm action association ratings from Binder et al. (2016)



2.3.5 Ease to pantomime

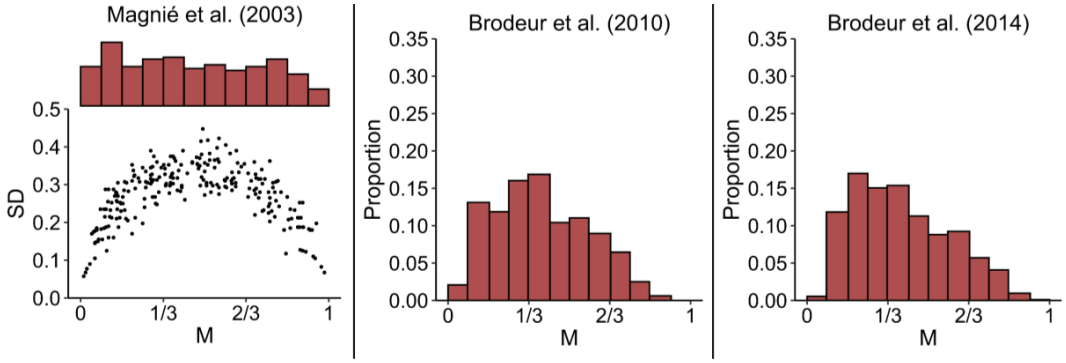
As discussed in the main text, two types of ease to pantomime instructions have been used in the literature. The first was introduced by Magnié et al. (2003) and assesses how easily the action associated to an object could be mimed so that someone else could recognise it. The second type of instructions are a modified version of the first and were proposed by Guérard et al. (2015). These do not include the part regarding the recognisability of the action and more generally assess how easily the use of an object could be mimed. The ratings resulting from the two instructions are analysed separately.

2.3.5.1 Original instructions

Similar to some of the previously reviewed instructions, the original ease to pantomime instructions appear to result in considerable disagreements (Figure B.31). In Brodeur et al. (2010) and Brodeur et al. (2014) in particular, respectively 47% and 43% of the items are in the middle of the scale, while close to none have consistently received high ratings. Magnié et al.'s

(2003) data displays a relatively more spread-out distribution. Note that this might be due to how their ratings were derived. In contrast to the two other studies – as well as to all the rating studies reviewed in the present work, responses to the scale's middle point (option 3 on a [1, 5] scale) were not included when computing the averages.

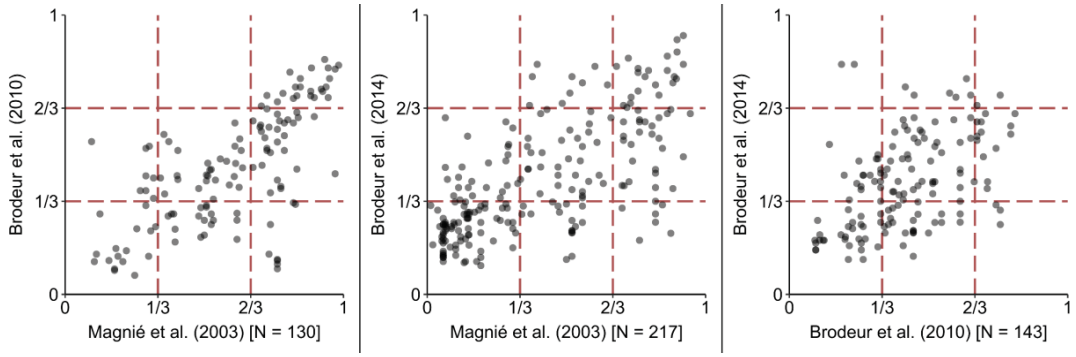
Figure B.31
Distributional histograms and the SDs against the means for the available ease to pantomime ratings (original instructions)



A comparison of the datasets (Figure B.32) suggests that the items in Magnié et al.'s (2003) study tended to generally receive higher ratings compared to Brodeur et al. (2010), and to a slightly lesser extent compared to Brodeur et al. (2014). We can indeed see that several items in the low or middle portions of the latter studies' scales were rated high in Magnié et al. (2003). Some of the low items in the two studies are also midscale in Magnié et al. (2003). Interestingly, a large portion of the items in all of these cases appear to be manipulable items such as tools, utensils and supplies (e.g. *axe*, *bottle*, *brush*, *hammer*, *knife*, *pen*, *pliers*, *spoon*). The comparison of the two datasets reported by Brodeur et al. (2010, 2014) reveals considerable variation for some items. This observation is surprising as both studies used virtually identical instructions, were conducted by the same main investigator, and included a comparable sample of participants. The rating procedures were nevertheless slightly different. Indeed, Brodeur et al. (2010) asked participants to rate the ease to pantomime in conjunction with other tasks and used an imposed rating time. On the other hand, the rating study of Brodeur et al. (2014) was self-paced and required participants to only rate the ease to pantomime.

Figure B.32

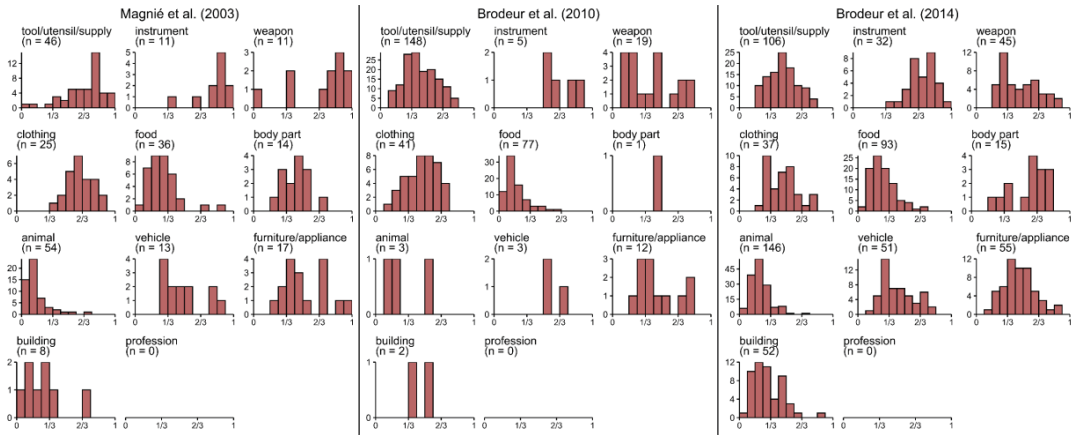
Comparison of the ease to pantomime ratings between datasets (original instructions)



The category-wise distributions also reveal important differences between the different datasets (Figure B.33). We can notably observe that typically FM items (tools, utensils, supplies, instruments and weapons) tend to be found on the high end of the scale in Magnié et al. (2003), whereas they have generated significant disagreements in Brodeur et al. (2010, 2014). Conversely, body parts are among the highest rated items in Brodeur et al. (2014) but are found midscale in Magnié et al. (2003). The other categories mostly display similar distribution patterns. For instance, clothes, furniture and appliances appear to generally be in the middle of the scale. Foods and animals are instead found towards the low end of the scale, although foods appear to have led to some disparity in the judgements.

Figure B.33

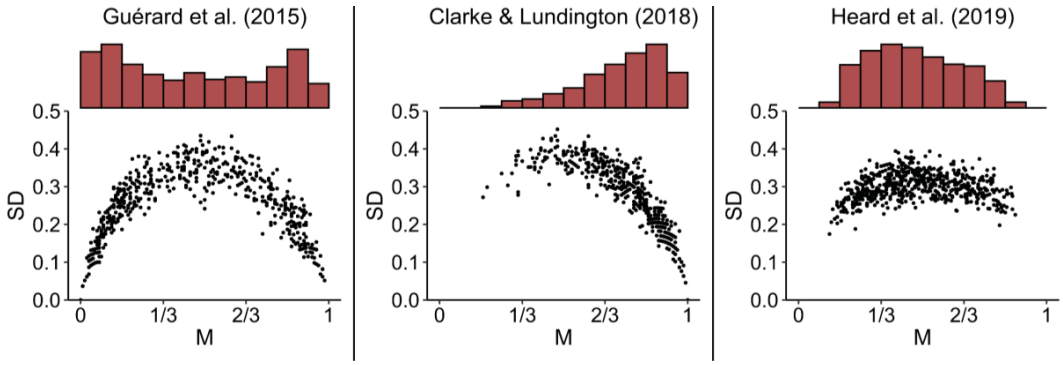
Category-wise rating distributions for the ease to pantomime ratings (original instructions)



2.3.5.2 Modified instructions

The ratings obtained through the modified instructions display distributions that differ considerably from those seen in the previous section – but also across the datasets that have collected them (Figure B.34). In Guérard et al.'s (2015) case, the ratings follow a roughly bimodal distribution with items on both the low and high ends of the scale. In contrast, they are heavily skewed toward the high end in Clarke and Lundington (2018) with no items on the low end of the scale. Finally, Heard et al.'s (2019) dataset mostly contains items in the middle of the scale (56%) or close, and no items rated consistently on the extremes.

Figure B.34
Distributional histograms and the SDs against the means for the available ease to pantomime ratings (modified instructions)

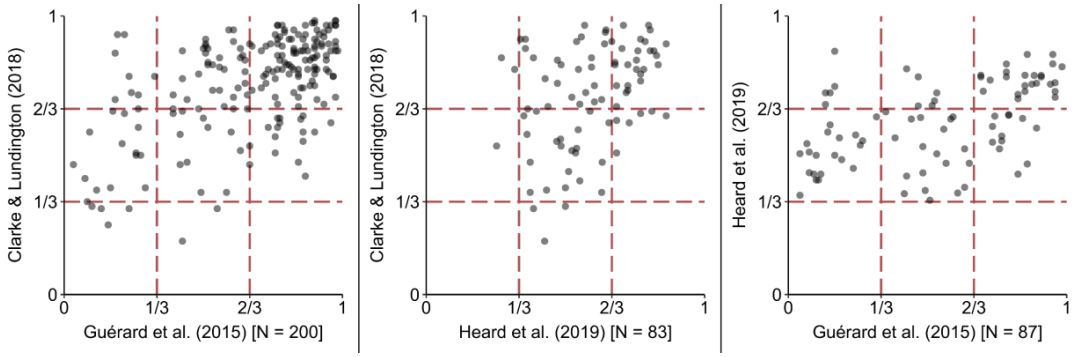


We have already raised some concerns in Section 2.2.1 about the reliability of Clarke and Lundington's (2018) and Heard et al.'s (2019) ease to grasp ratings. The comparison of the ease to pantomime ratings reported by these two studies to those of Guérard et al. (2015) largely support this initial analysis. We can see in Figure B.35 (left plot) that items rated low in Guérard et al. (2015) are mostly found midscale or on the high end of the scale in Clarke and Lundington (2018. E.g. *branch*, *picture frame*, *pylon*, *seashell*, *statue*). Some objects that can be reasonably expected to be rated high have also generated disagreements in the latter study, whereas they are rated high in Guérard et al. (2015. E.g. *brush*, *cigarette*, *corkscrew*, *dental floss*). On the other hand, Heard et al.'s (2019) study also shows some overlap with Guérard et al. (2015) on the high end of the scale (right plot). However, items both in the middle and in the low end of the scale in the latter study are either disagreed on or rated high in Heard et al. (2019). Finally, the comparison between Clarke and Lundington (2018) and Heard et al. (2019) is not very informative as most items were rated high in the first case and are in the middle of the scale in the latter (middle plot). Note that the correlation of the ease to pantomime ratings from these

two studies to all the other FM dimensions are also notably low or not statistically significant (Table B.11).

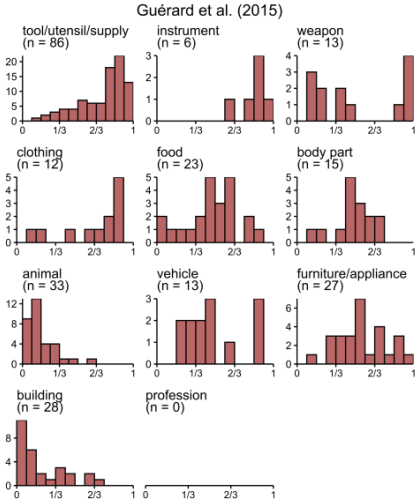
Figure B.35

Comparison of the ease to pantomime ratings between datasets (modified instructions)



In light of the above observations, analysing the category-wise distributions for Clarke and Lundington (2018) and Heard et al. (2019) would not be very informative. Figure B.36 thus only plots the distributions for Guérard et al.'s (2015) ratings. We can see that most tools, utensils, supplies and instruments are found on the high end of the scale. The first panel also displays some midscale items (e.g. *hook*, *jar*, *nail*, *rubber band*, *screw*, *tupperware*), but the scale appears to reliably capture most FM objects on its high end. The dataset contains very few weapons which appear on both ends of the scale. Among those on the low portion, *bow* and *shield* are the only ones that could be considered FM. Most clothing items are found on the high end of the scale, whereas animals and buildings are consistently rated high. Finally, the rest of the categories (i.e. foods, body parts, vehicles, furniture, appliances) appear to lead to disagreements.

Figure B.36
Category-wise rating distributions for the ease to pantomime ratings (modified instructions)

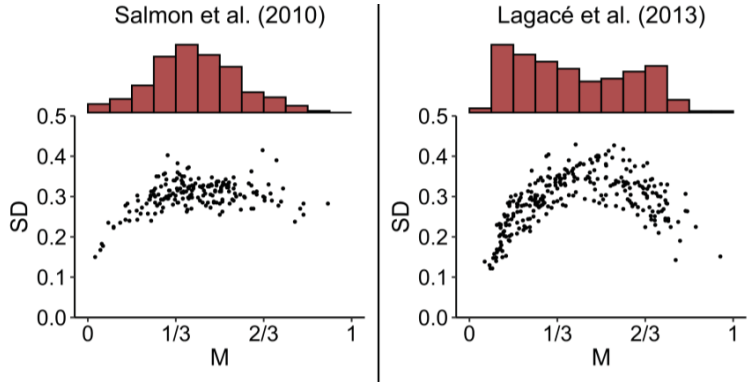


2.3.6 Grasp-use dissimilarity

The two studies reporting grasp-use dissimilarity ratings differ in their methodologies. Salmon et al. (2010) first asked their participants to rate the extent to which objects can be grasped and used. The grasp-use dissimilarity scale was only presented to them if their first rating was higher than the middle of the scale. Additionally, their instructions assessed the extent to which *hand movements* differ between picking objects up and using them. In contrast, participants in Lagacé et al.'s (2013) study rated all items for the difference in *hand posture* when grasping or using them.

As can be seen in Figure B.37, most items in Salmon et al.'s (2010) study can be found in the middle (60%) or toward the middle of the scale, with notably few items rated on the high end. Lagacé et al.'s (2013) ratings similarly contain very few high items. We can nevertheless observe relatively more items in the low end, and some items clustered in the higher end of the middle of the scale.

Figure B.37
Distributional histograms and the SDs against the means for the available grasp-use dissimilarity ratings



Although only a small number of items are in common between the two datasets, Figure B.38 suggests that Lagacé et al.'s (2013) instructions were more sensitive to the objects that had similar grasp and use postures.

Figure B.38
Comparison of the grasp-use dissimilarity ratings between datasets

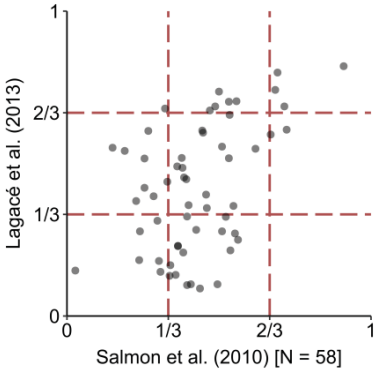
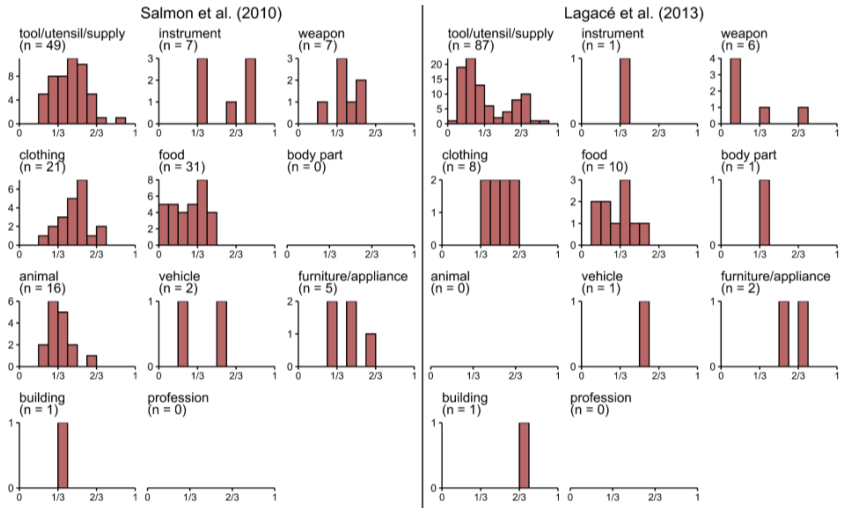


Figure B.39 shows that most categories in Salmon et al. (2010) are unsurprisingly distributed around the middle of the scale, although foods appear to have been rated relatively lower. Lagacé et al.'s (2013) dataset contains few – if any – items in some of the categories, which constrains our conclusions. We can nevertheless observe a bimodal distribution for tools, utensils and supplies. Some of these objects were mostly rated low on the scale (e.g. *axe*, *cane*, *cup*, *kettle*, *razor*), while another cluster of objects must have received high ratings from some participants (e.g. *calculator*, *dice*, *lighter*, *cell phone*). These results are difficult to interpret as

it is not entirely clear why the latter items were not consensually rated on the high end of the scale.

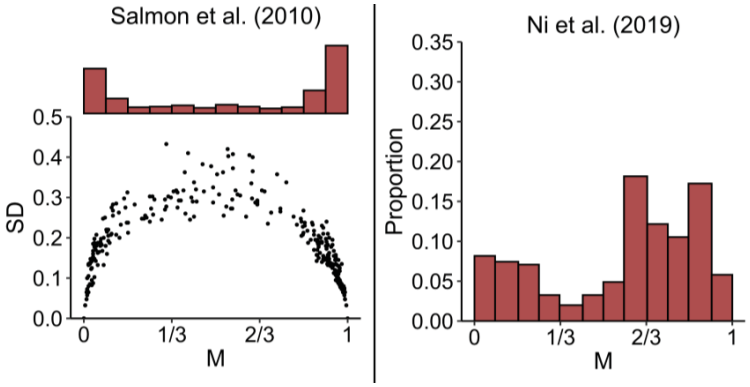
Figure B.39
Category-wise ratings distribution for the grasp-use dissimilarity ratings



2.4 Mixed manipulability

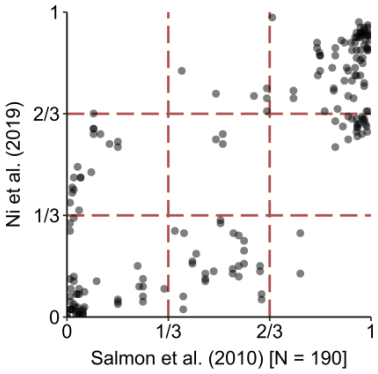
Two studies have collected mixed manipulability ratings and have used slightly different wordings. Salmon et al. (2010) asked participants to rate how easily objects could be grasped *and* used. Ni et al. (2019), on the other hand, assessed the extent to which they could be grasped *or* used. Despite this small difference, the distribution of ratings in the two studies are unexpectedly different (Figure B.40). Salmon et al.'s (2010) ratings are strongly clustered at the extreme ends of the scale. In contrast Ni et al.'s (2019) items appear to have received relatively more disparate ratings.

Figure B.40
Distributional histograms and the SDs against the means for the mixed manipulability ratings



Directly comparing the two datasets clearly reveals divergent response patterns (Figure B.41). We can notably see that almost no items are concurrently found in the middle of the scale: disagreed upon items in one study appear to be more strongly rated low or high in the other. For the midscale items in Ni et al. (2019), Salmon et al.'s (2010) study generally reports low ratings for vehicles, furniture and appliances (e.g. *airplane*, *bed*, *bus*, *car*, *desk*, *dresser*, *refrigerator*) and high ratings for foods (e.g. *carrot*, *cherry*, *lemon*, *onion*, *tomato*). Conversely, mostly animals (e.g. *bee*, *cat*, *frog*, *snake*, *turtle*) are disagreed on in Salmon et al. (2010) but were rated low in Ni et al. (2019).

Figure B.41
Comparison of the mixed manipulability ratings between datasets

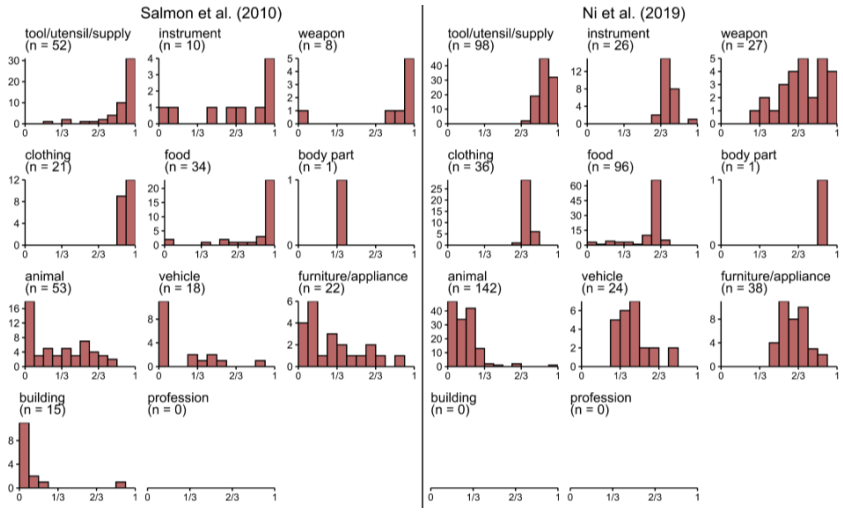


Category-wise analyses similarly display some differences between the two datasets. We can notably see that weapons, clothes and foods are found in or close to the middle of the scale in Ni et al. (2019), whereas they were more consistently rated high in Salmon et al. (2010).

Similarly, vehicles, furniture and appliances tend to have low ratings in the latter study but to generate disagreements in Ni et al. (2019). We can finally note that although most animals are found on the low end of the scale in both studies, several in Salmon et al. (2010) can also be found midscale.

Figure B.42

Category-wise rating distributions for the mixed manipulability ratings



Inspecting the correlations between the ratings from these two datasets and the arguably most reliable ratings for SM and FM provides further insights about what they capture. We can see in Table B.12 that Salmon et al.'s (2010) ratings correlate very strongly with SM variables (i.e. ease to grasp, likelihood to grasp, ease to hold). Ni et al.'s (2019) ratings also display high, but relatively lower correlations with these dimensions. The opposite pattern largely holds for FM variables (i.e. hand/arm action association, hand necessity for function, modified ease to pantomime). Indeed, Ni et al.'s (2019) ratings have relatively higher correlations with these variables than do those of Salmon et al. (2010). This result strongly suggests that Salmon et al.'s (2010) ratings capture SM despite mentioning the use of objects in their instructions. On the other hand, Ni et al.'s (2019) study appears to primarily capture FM.

Table B.12
Pair-wise Pearson correlation coefficients between each of the two mixed manipulability ratings and the most reliable SM and FM ratings

Study	Variable	Salmon et al. (2010)	Ni et al. (2019)
Guérard et al. (2015)	Ease to grasp	0.97	0.76
Díez-Álamo et al. (2018)	Likelihood to grasp	0.94	0.70
Stoinski et al. (2023)	Ease to hold	0.95	0.78
Binder et al. (2016)	Hand/arm action association	0.47	0.71
Moreno-Martínez & Montoro (2012)	Hand necessity for function	0.67	0.90
Miklashevsky (2018)	Hand necessity for function	0.62	0.88
Navarrete et al. (2019)	Hand necessity for function	0.62	0.86
Guérard et al. (2015)	Ease to pantomime (modified)	0.57	0.86

Note. All $p < .001$

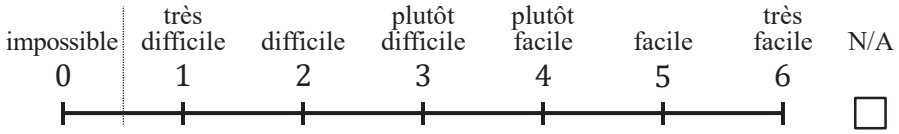
Appendix C
Supplementary materials for Chapter 5

1 Instructions for the graspability rating task

Veuillez évaluer la facilité avec laquelle vous pouvez saisir ce que représente chaque mot avec vos mains pour le bouger.

Exemples : Il est très facile de saisir et de bouger une TASSE, alors qu'il est très difficile de le faire avec une ARMOIRE. Quant à un AVION, celui-ci ne peut pas être manuellement saisi et bougé.

Vos évaluations devront être réalisées sur une échelle de 0 à 6, en choisissant la valeur que vous jugez appropriée à l'aide du curseur. Une valeur de 0 indique que ce que représente le mot est impossible à saisir avec les mains. Une valeur de 1 signifie que sa saisie est très difficile et une valeur de 6 que celle-ci est très facile. Les valeurs de 2 à 5 représentent des scores intermédiaires. Si vous ne connaissez pas un mot, et uniquement dans ce cas, vous pouvez cocher la case « N/A ».



Précisions :

Certains mots peuvent être ambigus. Il vous est demandé de les interpréter, quand c'est possible, comme des objets. Par exemple, le mot « orange » doit être interprété comme le fruit, et non la couleur. Veillez cependant à interpréter les mots uniquement selon leur définition courante (c.à.d. ni dans un sens familier, ni en argot).

Essayez d'être le plus précis possible dans vos évaluations, sans tout de même passer trop de temps sur chaque mot.

English translation

Please rate the ease with which you can grasp what each word represents with your hands in order to move it.

Examples: It is very easy to grasp and move a CUP, while it is very difficult to do so with a WARDROBE. As for an AIRPLANE, it cannot be manually grasped and moved.

Your ratings must be made on a scale from 0 to 6, by choosing the value you judge appropriate using the slider. A value of 0 indicates that what the word represents is impossible to grasp with your hands. A value of 1 means that grasping it is very difficult, and a value of 6 means that it is very easy. Values from 2 to 5 represent intermediate scores. If, and only if, you do not know a word, you may check the "N/A" box.

[Scale: 0 – impossible, 1 – very difficult, 2 – difficult, 3 – somewhat difficult, 4 – somewhat easy, 5 – easy, 6 – very easy, N/A]

Clarifications:
Some words can be ambiguous. You should interpret them, when possible, as objects. For example, the word "orange" should be interpreted as the fruit, not the colour. However, make sure to interpret the words only according to their common definition (i.e., not in a colloquial or slang sense).

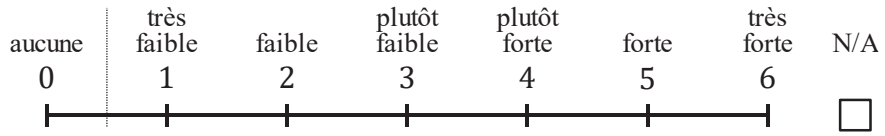
Try to be as precise as possible in your ratings, without spending too much time on each word.

2 Instructions for the action association rating task

Veuillez évaluer à quel point ce que représente chaque mot est associé à une utilisation avec les mains.

Exemples : Notre expérience de certains objets, comme un TIRE-BOUCHON ou un TORCHON, est intimement liée aux actions manuelles que nous réalisons pour les utiliser. Ces objets sont ainsi très fortement associés à une utilisation avec les mains. D'autres objets ne sont que faiblement associés à des actions manuelles pour leurs utilisations. Un TABLEAU, par exemple, peut être accroché, ou un VENTILATEUR peut être mis en marche, mais leurs usages sont ensuite indépendants d'une action manuelle. Finalement, certains objets ne sont associés à aucune utilisation avec les mains, comme un PLAFOND ou une COMÈTE.

Vos évaluations devront être réalisées sur une échelle de 0 à 6, en choisissant la valeur que vous jugez appropriée à l'aide du curseur. Une valeur de 0 indique que ce que représente le mot n'est associé à aucune utilisation manuelle. Une valeur de 1 signifie que l'association à une utilisation manuelle est très faible et une valeur de 6 qu'elle est très forte. Les valeurs 2 à 5 représentent des scores intermédiaires. Si vous ne connaissez pas un mot, et uniquement dans ce cas, vous pouvez cocher la case « N/A ».



Précisions :

Certains mots peuvent être ambigus. Il vous est demandé de les interpréter, quand c'est possible, comme des objets. Par exemple, le mot ORANGE doit être interprété comme le fruit, et non la couleur.

Certains mots peuvent vous faire penser à d'autres mots qui leurs sont associés (par ex. CLOU – MARTEAU). Il est important que vous jugiez uniquement l'utilisation manuelle de ce que représente le mot donné (CLOU), et non l'utilisation des mots associés (MARTEAU).

Essayez d'être le plus précis possible dans vos évaluations, sans tout de même passer trop de temps sur chaque mot.

English translation

Please rate the extent to which what each word represents is associated with manual use.

Examples: Our experience with certain objects, like a CORKSCREW or a DISHCLOTH, is closely tied to the manual actions that we perform to use them. These objects are thus strongly associated with a usage with the hands. Other objects are only weakly associated with manual actions in their usage. A PAINTING, for example, can be hung, or a FAN can be turned on, but their use is independent of manual actions afterwards. Finally, some objects are not associated with any use with the hands, such as a CEILING or a COMET.

Your ratings must be made on a scale from 0 to 6, by choosing the value you judge appropriate using the slider. A value of 0 indicates that what the word represents is not associated with any manual use. A value of 1 means that the association with manual use is very weak, and a value of 6 means it is very strong. Values from 2 to 5 represent intermediate scores. If, and only if, you do not know a word, you may check the "N/A" box.

[Scale: 0 – none, 1 – very weak, 2 – weak, 3 – somewhat weak, 4 – somewhat strong, 5 – strong, 6 – very strong, N/A]

Clarifications:
Some words can be ambiguous. You should interpret them, when possible, as objects. For example, the word "orange" should be interpreted as the fruit, not the colour.
Some words may remind you of other related words (e.g., NAIL – HAMMER). It is important that you only judge the manual use of the given word (NAIL), not the use of related words (HAMMER).

Try to be as precise as possible in your ratings, without spending too much time on each word.

3 Instructions for the functional attributes rating tasks

Usability

Est-il possible d'utiliser avec les mains ce que représente le mot ?

Oui / Non / Je ne connais pas ce mot

Number of actions

Veuillez estimer le nombre d'actions manuelles qui peuvent être typiquement réalisées avec l'objet.

Certains objets peuvent être utilisés pour réaliser plusieurs actions, alors que d'autres sont typiquement utilisés pour une action unique. Par exemple, un BÂTON peut être utilisé comme support, pour pousser ou rapprocher un autre objet, pour frapper, etc. En revanche, une seule action peut typiquement être réalisée avec un MARQUE-PAGE.



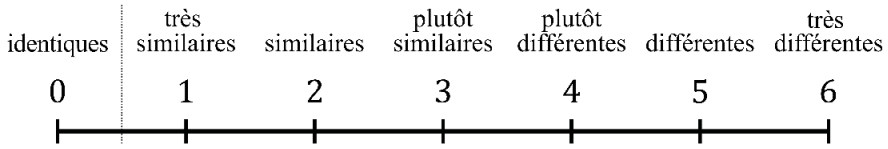
Précision:

Même si un objet peut être utilisé pour d'autres actions (par exemple, utiliser un couteau pour visser), nous vous demandons ici d'évaluer le nombre d'actions qui sont le plus couramment associées aux objets.

Move-use dissimilarity

Veuillez évaluer à quel point la posture des mains pour utiliser l'objet diffère de la posture nécessaire pour simplement le saisir et le bouger.

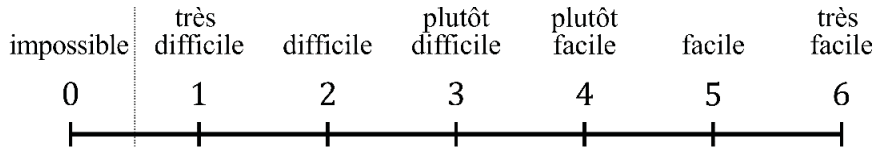
La posture des mains pour utiliser certains objets est très similaire à la posture nécessaire pour les transporter. Par exemple, une BROSSE À DENTS est saisie environ de la même façon pour son transport et pour son utilisation. La façon dont des BAGUETTES (chinoises) sont saisies, au contraire, diffère significativement selon si on veut les transporter ou les utiliser.



Pantomime

Imaginez que vous mimez l'utilisation typique de l'objet à une personne, sans lui parler. A quel point serait-il facile pour la personne de reconnaître l'objet dont il est question ?

Certains objets sont utilisés avec des actions manuelles qui leurs sont très spécifiques. Par exemple, l'action d'utilisation d'une TRONÇONNEUSE est très caractéristique de l'objet. Ainsi, il serait très facile pour une personne de le reconnaître en voyant l'action être mimée. En revanche, lorsqu'un grand nombre d'objets sont utilisés de façon similaire, il est très difficile pour une personne de reconnaître l'objet à partir du mime de son utilisation. Des BONBONS, par exemple, seraient très difficiles à reconnaître à partir du mime de l'action de les manger.



Précision:

Vos réponses doivent se baser sur le mime de l'**action pour utiliser l'objet,** et non sur un mime imitant l'objet (pour des ciseaux, par exemple, votre jugement doit porter sur l'action de tenir et d'utiliser l'objet et pas sur sa représentation avec les doigts).

English translation

Usability

Is it possible to use what the word represents with your hands?

Yes / No / I don't know this word

Number of actions

Please rate the number of manual actions that can typically be performed with the object.

Some objects can be used for multiple actions, while others are typically used for only one action. For example, a STICK can be used as support, to push or pull another object, to strike, etc. On the other hand, only one action is typically performed with a BOOKMARK.

[Scale: 0, 1, 2, 3, 4, 5, 6]

Clarification:
Even if an object can be used for other actions (e.g., using a knife as a screwdriver), we ask that you evaluate the number of actions most commonly associated with the objects.

Move-use dissimilarity

Please rate the extent to which the hand posture to use the object differs from the posture needed to simply grasp and move it.

The hand posture for using some objects is very similar to the posture needed to transport them. For example, a TOOTHBRUSH is held in about the same way for transport and for use. Conversely, the way CHOPSTICKS are held differs significantly depending on whether you are transporting or using them.

[Scale: 0 – identical, 1 – very similar, 2 - similar, 3 – somewhat similar, 4 – somewhat different, 5 – different, 6 – very different]

Pantomime

Imagine that you are miming the typical use of the object to someone, without speaking. How easy would it be for the person to recognize the object?

Some objects are used with very specific manual actions. For example, the use of a CHAINSAW is very characteristic of the object. Thus, it would be very easy for someone to recognize it by seeing the action being mimed. On the other hand, when many objects are used in a similar way, it is very difficult for someone to recognize the object based on its mime. CANDY, for example, would be very difficult to recognize from miming the action of eating it.

[Scale: 0 – impossible, 1 – very difficult, 2 – difficult, 3 – somewhat difficult, 4 – somewhat easy, 5 – easy, 6 – very easy, N/A]

Clarification:
Your answers should be based on miming the **action to use the object**, not the object itself (for scissors, for example, your judgment should be based on the action of holding and using the object, not on representing it with your fingers).

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Université de Toulouse

Titre : De la manipulabilité à la méthodologie et vice-versa : un regard critique sur les pratiques expérimentales dans l'étude des représentations d'objets

Mots clés : Normes sémantiques, Echelle de Likert, Manipulabilité, Affordances, Validité méthodologique, Cognition incarnée

Résumé : L'idée que la cognition est incarnée et que les systèmes sensori-moteurs jouent un rôle central dans ses différents processus a eu un impact majeur sur notre compréhension actuelle de l'organisation et de la représentation des connaissances. Bien qu'elle ait gagné un soutien important au fil des années, cette approche a également suscité de nombreux débats et critiques. Une partie des désaccords provient notamment de résultats empiriques qui semblent difficiles à concilier avec certaines de ses prémices centrales. Cette thèse se focalise sur une partie de cette littérature qui est fréquemment citée en faveur d'un ancrage multimodal des connaissances, à savoir les études sur le rôle des informations motrices dans la reconnaissance et le traitement des objets manipulables. A travers une revue étendue et détaillée, nous soutenons que les résultats sur ce sujet restent largement non concluants, en partie en raison de problèmes de validité méthodologique. L'objectif de cette thèse est de remettre en question une pratique répandue en particulier : l'utilisation de normes de type Likert (variables subjectives) pour la sélection et le contrôle des stimuli. Nous commençons par une investigation générale de la fiabilité et de la validité des variables subjectives, et présentons un certain nombre d'implications importantes pour les études qui les utilisent. À la lumière de ce premier travail, nous nous penchons ensuite sur l'opérationnalisation de la manipulabilité des objets et proposons une analyse et une discussion approfondies de ses diverses définitions dans la littérature. Nous présentons enfin de nouvelles normes de manipulabilité pour des mots français qui nous permettent de comparer directement les résultats obtenus à travers différentes consignes, et ainsi de complémenter nos analyses sur leurs validités respectives. Nos résultats révèlent globalement une utilisation très flexible et souvent inappropriée des variables subjectives, ce qui complique les comparaisons entre les études et affecte directement la validité des expériences. Ils fournissent en outre un éclairage essentiel sur les propriétés des évaluations de type Likert, nous permettant ainsi de proposer des recommandations pour des pratiques plus robustes et des pistes pour de futures recherches.

Title: From manipulability to methodology and back: a critical look into experimental practices in the study of object representations

Key words: Semantic ratings, Likert scale, Manipulability, Affordances, Methodological validity, Embodied cognition

Abstract: The view that cognition is embodied and that sensory-motor systems play a central role in its various processes has had a major impact on our current understanding of how knowledge is organised and represented. Despite having gathered substantial support over the years, this account has nevertheless been extensively debated and criticised. Some of the disagreements notably stem from empirical findings that appear difficult to reconcile with some of its core premises. The current work focuses on a subset of this literature that is frequently cited in support of multimodal knowledge representations, namely on studies investigating the role of motor information in the recognition and processing of manipulable objects. Through a broad and critical review, we argue that the evidence on this topic remains largely inconclusive, partly due to issues with methodological validity. The goal of this thesis is to question one widespread practice in particular: the use of Likert-type scale ratings (subjective variables) for stimulus selection and control. We start with a general investigation of the reliability and validity of subjective variables, and lay out a number of important implications for the studies using them. Armed with new insights about what such ratings represent, we then turn to how object manipulability has been operationally defined in the literature and offer an in-depth analysis and discussion of its different assessments. Finally, we present a new set of manipulability ratings for French words that we use to directly compare the results obtained through different rating instructions and to discuss their respective validities further. Overall, our results reveal a highly flexible and often inappropriate use of subjective variables, which both complicates cross-study comparisons and directly affect the validity of experiments. They additionally provide critical insights into the properties of Likert-type ratings, allowing us to propose recommendations for more robust practices and avenues for further inquiry.