



HAL
open science

Addressing the Elephant in the Middle: Implications of the Midscale Disagreement Problem Through the Lens of Body-Object Interaction Ratings

Dimitri Paisios, Nathalie Huet, Elodie Labeye

► **To cite this version:**

Dimitri Paisios, Nathalie Huet, Elodie Labeye. Addressing the Elephant in the Middle: Implications of the Midscale Disagreement Problem Through the Lens of Body-Object Interaction Ratings. *Collabra: Psychology*, 2023, 9 (1), 10.1525/collabra.84564 . hal-04405485

HAL Id: hal-04405485

<https://hal.science/hal-04405485v1>

Submitted on 19 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.



L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Methodology and Research Practice

Addressing the Elephant in the Middle: Implications of the Midscale Disagreement Problem Through the Lens of Body-Object Interaction Ratings

Dimitri Paisios¹, Nathalie Huet¹, Elodie Labeye¹

¹ Laboratoire Cognition, Langues, Langage, Ergonomie, University of Toulouse Jean Jaurès, Toulouse, France

Keywords: Likert scale, semantic ratings, midscale disagreement, body-object interaction

<https://doi.org/10.1525/collabra.84564>

Collabra: Psychology

Vol. 9, Issue 1, 2023

When participants disagree about their judgments on a Likert-type scale, the average rating will be naturally drawn towards its middle. The present work's goal is to explore the implications of this *midscale disagreement problem* for psycholinguistic norms by using the literature on Body-Object Interaction (BOI) ratings as a case study. Through a series of graphical analyses, we argue that (i) the average rating of most midscale items cannot be interpreted as their true position on the variable's continuum; (ii) other variables driving the disagreement in judgements can introduce an independent midscale effect in word processing performances; (iii) the typical sample sizes used by norming studies are likely insufficient to reliably detect disagreements and can lead to significant measurement error. A methodological review of the studies on BOI's effect in word processing reveals that most of them suffer from the midscale disagreement problem, either because of inadequate word sampling or statistical modelling. Whereas these observations provide initial clues for the interpretation and use of the ratings, it remains difficult to determine the full scope of the disagreement problem based only on the summary statistics reported by rating studies. To address this point, we present new BOI ratings for a set of 1019 French words which we use to perform item-level descriptive and exploratory analyses. Overall, the results confirm that unipolar Likert-type scale ratings such as BOI capture the dimension of interest mainly at the two ends of the scale, while they represent increasing disagreement among participants as they approach the middle. These observations provide initial best-practice recommendations for the use and interpretation of subjective variables. Our analyses can additionally serve as general guidelines to interpret similar ratings and to assess the validity of previous findings in the literature based on standard summary statistics.

The idea that concepts are grounded in the same systems through which they are acquired (Barsalou, 1999, 2008) has attracted a lot of interest in the role of sensorimotor information in lexical-semantic processing. The resulting need for studies to test and to control for the experiential attributes associated to stimuli has translated into a recent proliferation of subjective (also called semantic) norms. To name just a few, ratings have been collected for sensory experience (Bonin et al., 2015; Juhasz & Yap, 2013), general and modality-specific perceptual and action strengths (Chedid et al., 2019; Lynott et al., 2020; Miceli et al., 2021; Speed & Majid, 2017; Vergallito et al., 2020) and various manipulation-related dimensions (e.g. graspability, ease of pantomime, number of actions; Amsel et al., 2012; Heard

et al., 2019). Along with megastudies providing word processing performances (e.g. Balota et al., 2007; Ferrand et al., 2018), the increasing availability of such norms plays a crucial role in advancing our understanding of how words are recognised and how their meanings emerge.

Despite their importance, the ratings in themselves have been subject to surprisingly little methodological consideration in the psycholinguistics literature. Norming studies typically ask participants to subjectively rate a list of words 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 the item's position on a continuum. As Pollock (2018) recently pointed out, this introduces an important confound which has been

a Correspondence concerning this article should be addressed to Dimitri Paisios, Laboratoire Cognition, Langues, Langage, Ergonomie, Maison de la Recherche, University of Toulouse Jean Jaurès, 5 Allées Antonio Machado, 31058, Toulouse, France. Email: dimitripaisios@gmail.com

almost entirely overlooked. If participants disagree in their judgements for a word (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. As will be discussed below, this *midscale disagreement problem* has serious implications for studies investigating the effects of subjective variables and can lead to false inferences if it is not accounted for.

The current work offers an analysis of these implications through a case study based on the *body-object interaction* (BOI – Siakaluk, Pexman, Aguilera, et al., 2008) literature. After a brief overview of the variable's effect in word processing, we use available BOI norming datasets to outline three major consequences stemming from the midscale disagreement problem and explore the extent to which they affect the experiments on the variable's role in word processing. We additionally report new BOI ratings for a set of 1019 French words and the results of item-level descriptive and exploratory analyses which provide a more detailed look into the disagreement problem.

The Body-Object Interaction Effect

BOI captures the extent to which words' referents afford physical – and particularly manual (Heard et al., 2019) – interaction for the human body (Siakaluk, Pexman, Aguilera, et al., 2008). It is rated on a 7-point Likert-type scale, with higher values representing easier interactions, and has been interpreted as a general indicator of a concept's sensorimotor richness. A large number of studies have shown that words whose referents are easy to interact with (high BOI words, e.g. *chair*, *screwdriver*) are processed faster and more accurately than those for which it is harder to do so (low BOI words, e.g. *song*, *cloud*) in both lexical (Bennett et al., 2011; Hansen et al., 2012; Siakaluk, Pexman, Aguilera, et al., 2008; Tillotson et al., 2008; Yap et al., 2012) and semantic tasks (Bennett et al., 2011; Hansen et al., 2012; Hargreaves et al., 2012; Heard et al., 2019; Muraki et al., 2023; Muraki & Pexman, 2021; Pexman et al., 2019; Siakaluk, Pexman, Sears, et al., 2008; Wellsby et al., 2011; Yap et al., 2012), as well as in sentence reading (Phillips et al., 2012; Xue et al., 2015). To a large extent, these results have been taken as evidence that sensorimotor attributes are constitutive of semantic representations and that richer sensorimotor information facilitates lexical-semantic processing (Pexman, 2012, 2020). Additionally, BOI's effect has been shown to vary across different semantic decision requirements (Al-Azary et al., 2022; Newcombe et al., 2012; Tounsignant & Pexman, 2012. See also Muraki et al., 2023; Taikh et al., 2015), thus suggesting that sensorimotor information is flexibly drawn upon with respect to its relevance in a given context (Lebois et al., 2015; Yee & Thompson-Schill, 2016).

An open question remains, however, as to BOI's role in word recognition (see also Connell & Lynott, 2016). In contrast to the experiments reported above, several studies have failed to replicate the facilitatory BOI effect in lexical decision task (LDT) performances (Alonso et al., 2018; Hargreaves & Pexman, 2014; Heard et al., 2019; Juhasz et al.,

2011; Muraki & Pexman, 2021; Pexman et al., 2019). Connell and Lynott (2016) argued in their review 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 reporting a BOI effect. Some regression studies have nonetheless found a significant BOI effect over and above AoA – although with rather counterintuitive results. Bennett et al. (2011) included several lexical variables, AoA and imageability in their analysis and reported a facilitatory BOI effect on LDT latencies. Alonso et al. (2018), on the other hand, used practically the same model and found an inverse, inhibitory effect. Two studies using slightly different models found no effect of the variable above AoA (Heard et al., 2019; Juhasz et al., 2011). Most interestingly, the largest regression analysis to date by Pexman et al. (2019) over 3591 nouns revealed a quadratic BOI effect on LDT performances after controlling for lexical variables, concreteness and AoA. Latencies were shorter (and accuracies higher) for midscale words compared to those at the two ends of the scale, with no clear differences between low- and high-BOI words. These conflicting findings, and particularly the processing advantage found for words with moderate physical interaction ratings (midscale), are difficult to reconcile with any theoretical account of the variable. As will be argued below, they are likely due to methodological problems inherent to Likert-type ratings which affect a large number of studies on the subject.

The Midscale Disagreement Problem

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 1). 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 significant disagreement in the raters' judgments. For reference, a completely uniform response distribution on a 7-point 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

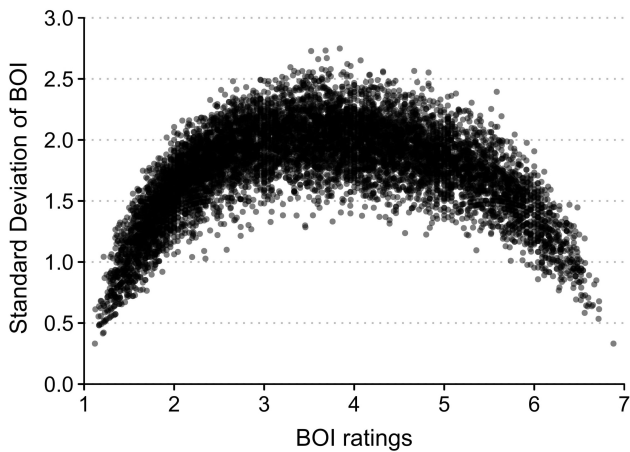


Figure 1. Standard deviations of the BOI ratings as a function of their means for 9351 words collected by Pexman et al. (2019)

the variable's continuum and cannot be interpreted as representing their position on the scale.

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 process-

ing 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 2.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 (2016) about a confound between AoA and BOI, Figure 2.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.

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

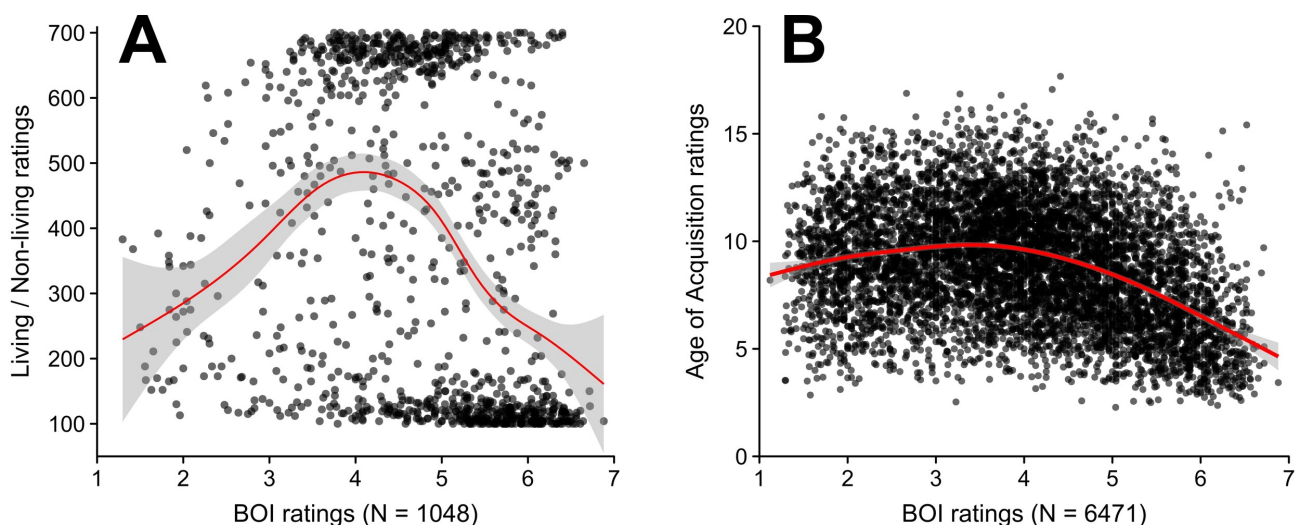


Figure 2. Living/Non-living (A; VanArsdall & Blunt, 2022) and age of acquisition (B; Kuperman et al., 2012) ratings against BOI ratings (Pexman et al., 2019)

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.

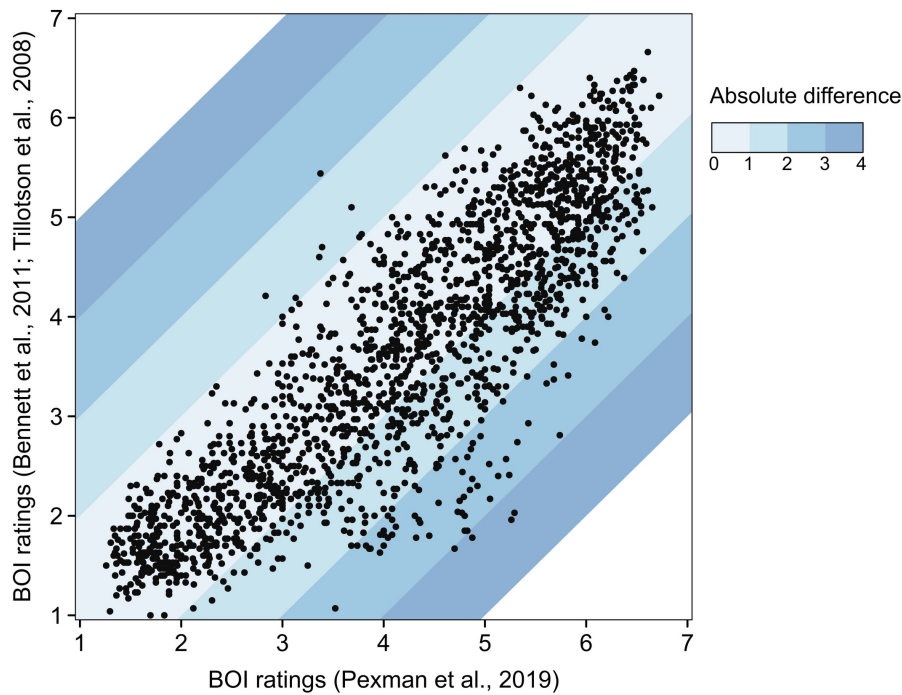


Figure 3. Combined BOI ratings from Bennett et al. (2011) and Tillotson et al. (2008) against those provided by Pexman et al. (2019) for 1897 words in common

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

the problem. Figure 3 plots the combined ratings from Bennett et al. (2011) and Tillotson et al. (2008)¹ 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 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 30 observations by word might be insufficient to obtain accurate rating distributions.

The following sections offer a methodological review of the experiments on BOI's effect in light of the issues outlined above. For clarity purposes, factorial design studies are reviewed separately from those using regression analyses.

Low vs. High BOI

Stimulus lists in factorial experiments are obtained through a high-low split (or dichotomisation) of the vari-

able 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 the high-BOI words were generally drawn from the high end of the scale. However, the low-BOI lists have averages close to the middle of the scale and were thus likely composed of a significant amount of midscale words. Two exceptions are Al-Azary et al.'s (2022) and Muraki and Pexman's (2021) recent experiments in which low-BOI words appear to have relatively lower ratings (but note the higher SDs for high-BOI word in the latter case).

Thanks to most of these studies sharing their stimulus sets, 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 4 depicts the stimuli used by Hargreaves et al. (2012) and Muraki and Pexman (2021) in their lexical decision tasks, plotted on top of all the words in their reference datasets. In the former case, most words in the low-BOI list are indeed 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 the other studies (except for Al-Azary et al., 2022; Muraki & Pexman, 2021)

¹ 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.

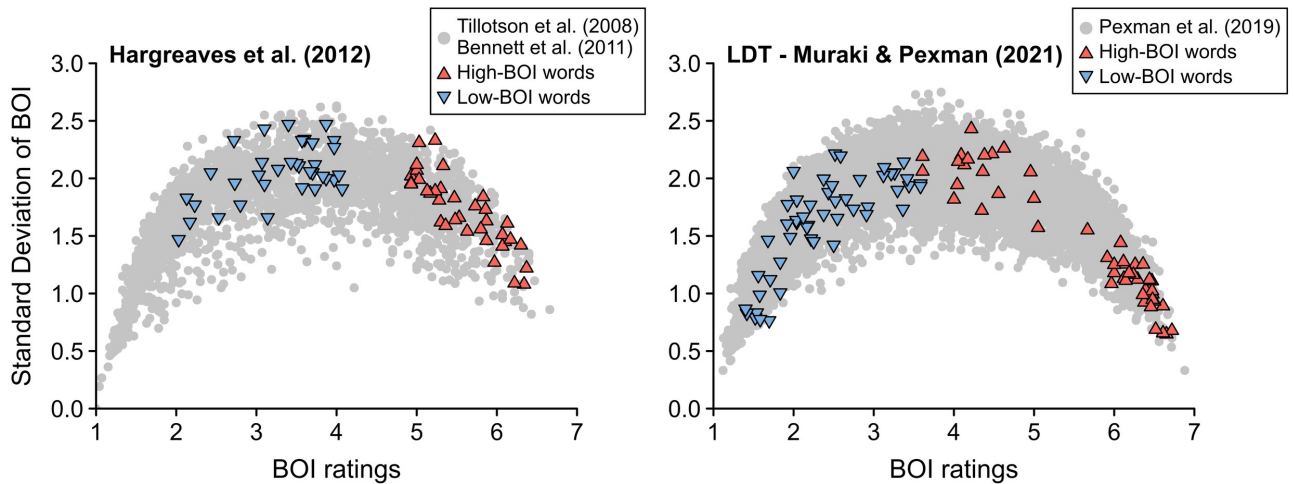


Figure 4. BOI ratings of the stimuli used by Hargreaves et al. (2012) and in the lexical decision task (LDT) from Muraki and Pexman (2021) against all the items in their reference datasets

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			
Siakaluk et al. (2008)	L decision	24	3.3	5.3
Siakaluk et al. (2008)	S & S-L decision			
Xue et al. (2015)	Sentence acceptability			
Al-Azary et al. (2022)	S decision	45	2.56 (.56)	5.55 (.38)
Hargreaves et al. (2012)	L decision	36	3.30 (.59)	5.60 (.47)
Muraki et al. (2023)	S decision	50	3.00 (.77)	5.59 (.44)
Muraki & Pexman (2021)	L decision	50	2.39 (.64)	5.58 (1.01)
Muraki & Pexman (2021)	Syntactic classification	50	2.04 (.47)	5.14 (1.04)
Phillips et al. (2012)	Sentence reading	40	3.33 (.59)	5.63 (.44)
Tousignant & Pexman (2012)	S decision	35	3.39 (.55)	5.67 (.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.

based on their summary statistics.² 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*. Muraki and Pexman (2021) used a different set of words with relatively more representative ratings of the entirety of the scale. These were nevertheless not clearly split and several display high disagreement which likely introduces considerable noise to the results.

Examining the between-study rating variability for these stimulus sets raises further concerns about factorial exper-

iments' validity. As can be seen in Figure 5, 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. This is particularly the case for the experiments which have used the Siakaluk et al. (2008) norms (Hansen et al., 2012; Siakaluk, Pexman, Aguilera, et al., 2008; Siakaluk, Pexman, Sears, et al., 2008; Wellsby et al., 2011; Xue et al., 2015), as well as the stimulus lists of Muraki and Pexman (2021) – although only a small portion of their words were available in the other datasets. Al-Azary et al.'s (2022) study is, to our knowledge, the only one

² The studies by Hansen et al. (2012), Siakaluk et al. (2008; Siakaluk, Pexman, Sears, et al., 2008), Xue et al. (2015) and Wellsby et al. (2011) used the ratings collected by Siakaluk et al. (2008) which are not publicly available. Muraki et al.'s (2023) 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).

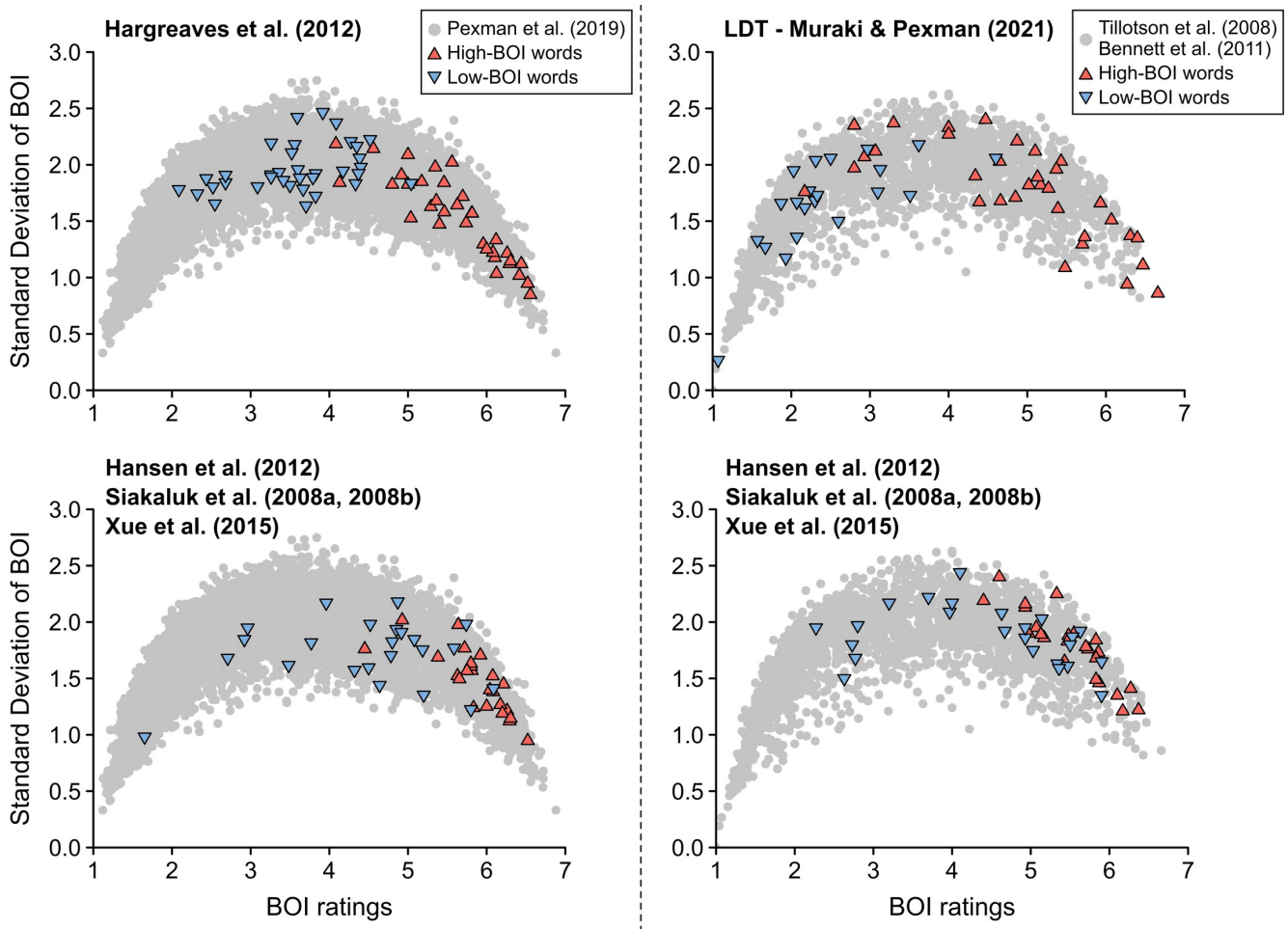


Figure 5. BOI ratings of the stimuli used by Hargreaves et al. (2012), the lexical decision task (LDT) in Muraki and Pexman (2021) and by Siakaluk et al. (2008a, 2008b; Hansen et al., 2012; Xue et al., 2015) as found in a different dataset than that which was originally used (left column: Pexman et al., 2019; right column: combined ratings from Bennett et al., 2011, and Tillotson et al., 2008)

in which the two lists display a healthy range of ratings in the datasets from which they were drawn (Bennett et al., 2011; Tillotson et al., 2008). In Pexman et al.'s (2019) ratings, however, the words in their low-BOI list are primarily found at the middle of the scale and mostly towards its high end ($M = 4.45$, $SD = .66$, $Min = 2.08$, $Max = 5.74$). 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. It is thus overall difficult to determine what these experiments ultimately compare and their results cannot be reliably attributed to an effect of BOI.

Regression Studies

Regression analyses have been described as a statistically much more robust method to assess a variable's in-

fluence in word processing (Balota et al., 2012; Brysbaert et al., 2014, 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 (Bennett 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 6³), 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,

³ It should be noted that the ratings presented in Figure 6 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.

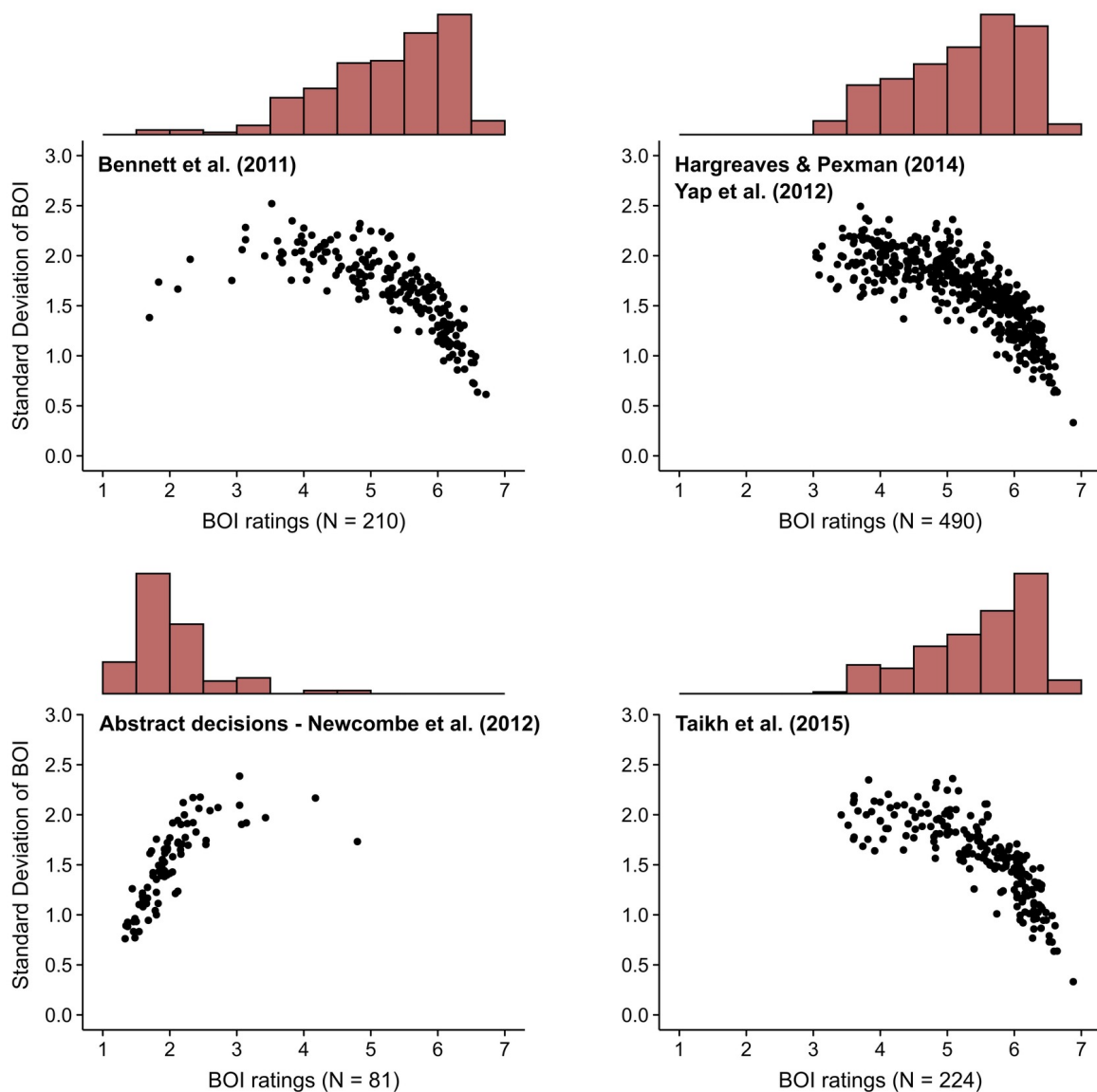


Figure 6. BOI ratings (Pexman et al., 2019) of the words included in the analyses of Bennett et al. (2011), Hargreaves & Pexman (2014; Yap et al., 2012), Newcombe et al. (2012) and Taikh et al. (2015).

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

no reliable inferences can thus be made about BOI's role in word processing based on these analyses.

To our knowledge, only six regression analyses 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; Newcombe et al., 2012; Pexman et al., 2019; Tillotson et al., 2008). With the exception of Pexman et al. (2019), 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, midscale words are susceptible to exhibit differences in processing relative to those at the ends of the scale due to confound 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.

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) as the analyses are 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 integrity of 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. 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.

The Present Study

The issues raised here and by Pollock (2018) are fundamentally based on the observation that words with mid-scale 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.

Method

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.

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 Supplemental Materials) 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.⁴ 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 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

⁴ 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.

Table 2. Packages used for data analyses

Package	Version	Authors
car	3.1-0	Fox & Weisberg (2019)
DescTools	0.99.45	Signorell et al. (2022)
ggthemes	4.2.4	Arnold (2021)
here	1.0.1	Müller (2020)
IDPmisc	1.1.20	Locher (2020)
paletteer	1.4.0	Hvitfeldt (2021)
psych	2.1.3	Revelle (2021)
readxl	1.4.2	Wickham & Bryan (2023)
tidyverse	2.0.0	Wickham et al. (2019)

reached were incrementally removed from the sampling pool.

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.

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). The packages used for the present work are presented in [Table 2](#). 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>).

Results

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 ± 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 ($SD = 2.84$, $Min = 75$, $Max = 96$).

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 3](#)). 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 correlations ($N = 28$, Alonso et al., 2018; $N = 7$, Bennett et

Table 3. 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.

Table 4. 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

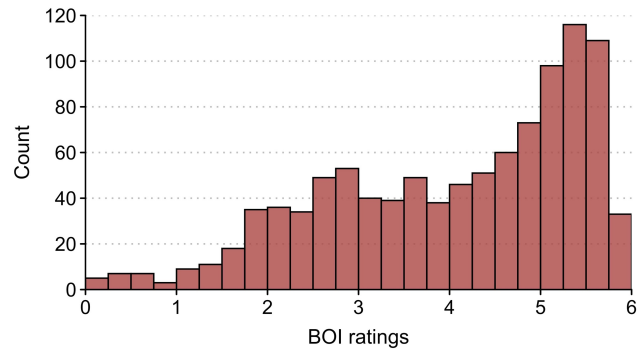
Note. All $ps < .001$.

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).

Descriptive Analysis

Descriptive statistics for the combined ratings are summarised in Table 5 and their frequency distribution is presented in Figure 7.

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 8. 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*, $M = 2.89$, $SD = 1.75$). 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

**Figure 7. Histogram of the present BOI ratings (N = 1019)**

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 depends 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 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 8). Similarly, the average rating be-

Table 5. 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	1 st Q	Med	3 rd 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

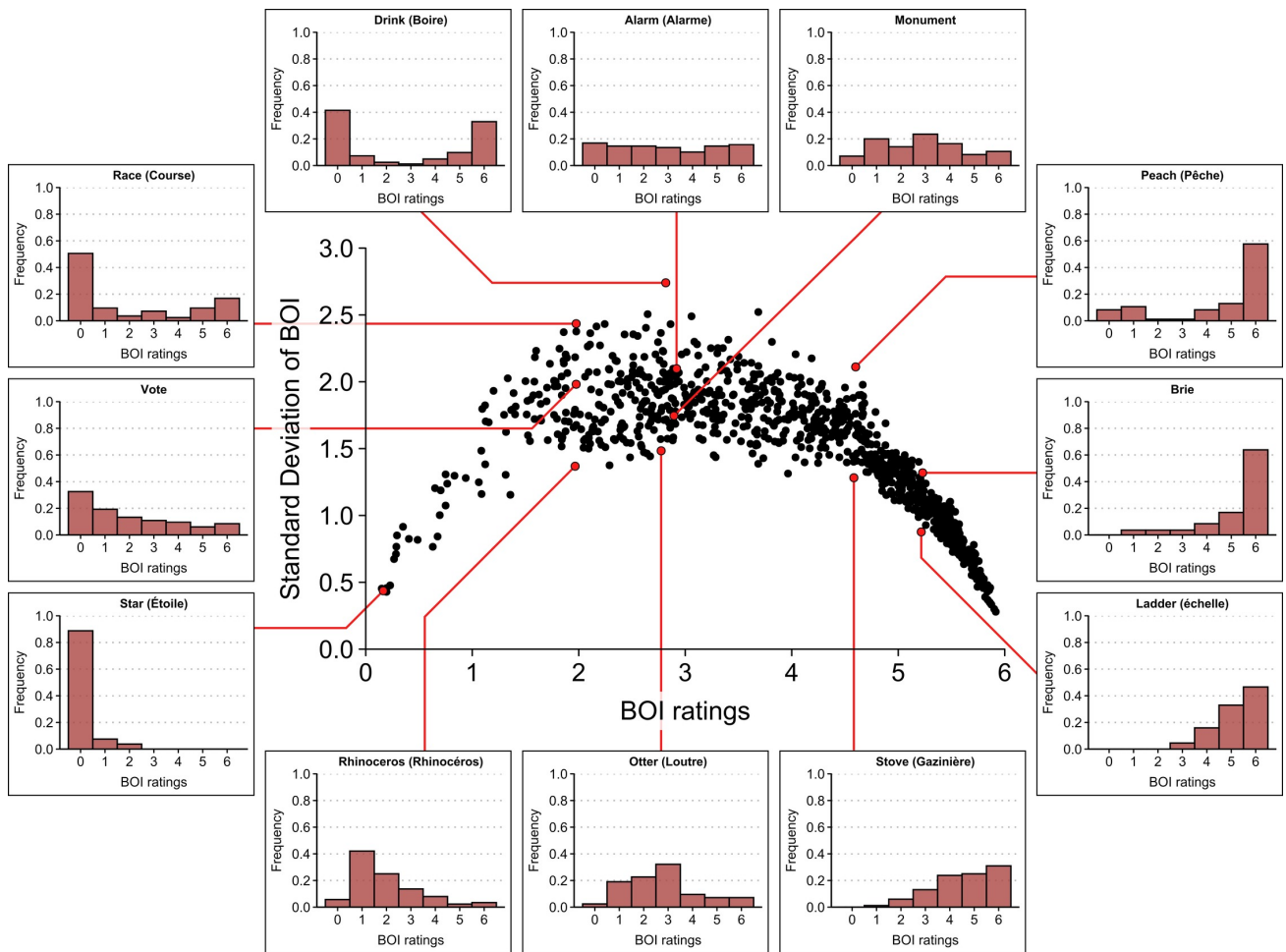


Figure 8. Standard deviation as a function of the average BOI ratings in the present study (centre) and examples of item-level rating distributions

comes 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.

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 Abdal Rahem & Darrah, 2018; Claveria, 2021; 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 5. 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 9 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 .65⁵ 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

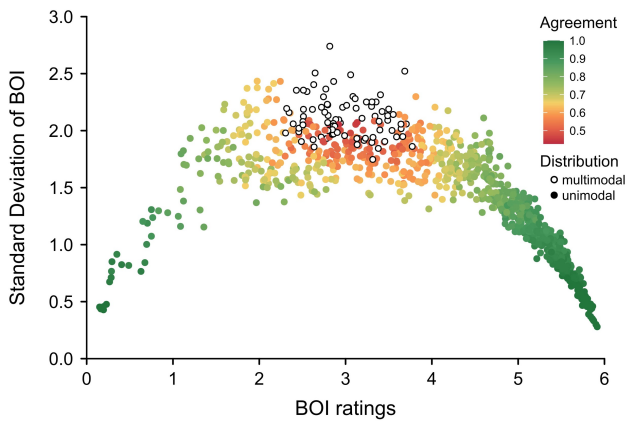


Figure 9. Standard deviation as a function of the average BOI ratings, along with item-level interrater agreement scores and type of rating distribution

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*, $M = 2.40$, $SD = 1.67$, $Agreement = .74$).

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.

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 5](#). [Figure 10](#) 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 ap-

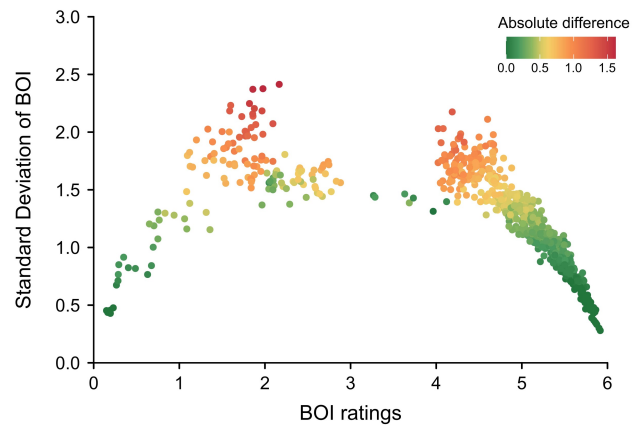


Figure 10. Standard deviation as a function of the average BOI ratings for words with an agreement score above .65 and absolute differences between the trimmed means and the average ratings

proximately 1.5, the trimmed mean is found either within [0, 1] or [5, 6].

Discussion

The present work’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 (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 Mon-

⁵ 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.

tamedi 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. Almost half of them 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), 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 see Bonin et al., 2018; Kousta et al., 2011). Their results should thus be interpreted with caution. It is important to note 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 and the reliability of its results. 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., 2023). 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. Note, however, that their small number 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. The analyses presented in the current work 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 valid-

ity 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.

Contributions

Contributed to conception and design: DP, NH, EL
 Contributed to acquisition of data: DP
 Contributed to analysis and interpretation of data: DP
 Drafted and/or revised the article: DP, NH, EL
 Approved the submitted version for publication: DP, NH, EL

Acknowledgements

We would like to thank Maëlys Home for her assistance in participant recruitment and data collection for the print version of the questionnaire, and all participants who volunteered to provide BOI ratings. We also thank the three anonymous reviewers for their comments on a previous version of this work. Their valuable insights helped us to significantly improve the manuscript and to explore the midscale disagreement problem's implications in much more detail.

Competing Interests

None of the authors has competing interests to disclose.

Data Accessibility Statement

All rating data (raw, pre-processed and summary) and the R script used for the analyses are openly available at the Open Science Framework repository associated to this article, at <https://osf.io/9murh>.

Supplemental Material

Body-Object Interaction Rating Task Instructions.

Submitted: June 07, 2023 PDT, Accepted: June 26, 2023 PDT



References

- Abdal Rahem, M., & Darrah, M. (2018). Using a Computational Approach for Generalizing a Consensus Measure to Likert Scales of Any Size. *International Journal of Mathematics and Mathematical Sciences*, 2018, e5726436. <https://doi.org/10.1155/2018/5726436>
- Akiyama, Y., Nolan, J., Darrah, M., Abdal Rahem, M., & Wang, L. (2016). A method for measuring consensus within groups: An index of disagreement via conditional probability. *Information Sciences*, 345, 116–128. <https://doi.org/10.1016/j.ins.2016.01.052>
- 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>
- 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>
- Arnold, J. B. (2021). *ggthemes: Extra Themes, Scales and Geoms for "ggplot2."* R package version 4.2.4. <https://cran.r-project.org/package=ggthemes>
- Balota, D. A., Yap, M. J., & Hutchison, K. A. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing?: David A. Balota, Melvin J. Yap, Keith A. Hutchison, and Michael J. Cortese. In *Visual Word Recognition* (Vol. 1). Psychology Press.
- 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>
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4), 577–660. <https://doi.org/10.1017/s0140525x99002149>
- Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, 59(1), 617–645. <https://doi.org/10.1146/annurev.psych.59.103006.093639>
- 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>
- 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., 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., & Bugajska, 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., Méot, A., Ferrand, L., & Bugajska, A. (2015). Sensory experience ratings (SERs) for 1,659 French words: Relationships with other psycholinguistic variables and visual word recognition. *Behavior Research Methods*, 47(3), 813–825. <https://doi.org/10.3758/s13428-014-0503-x>
- Bonin, P., Méot, A., Ferrand, L., & Roux, S. (2011). L'imageabilité: Normes et relations avec d'autres variables psycholinguistiques. *L'Année psychologique*, 111(2), 327–357. <https://doi.org/10.3917/anpsy.112.0327>
- 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>
- Brysbaert, M., Keuleers, E., & Mandera, P. (2014). A plea for more interactions between psycholinguistics and natural language processing research. *Computational Linguistics in the Netherlands Journal*, 4, 209–222.
- 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>
- 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>
- 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>
- Chedid, G., Brambati, S. M., Bedetti, C., Rey, A. E., Wilson, M. A., & Vallet, G. T. (2019). Visual and auditory perceptual strength norms for 3,596 French nouns and their relationship with other psycholinguistic variables. *Behavior Research Methods*, 51(5), 2094–2105. <https://doi.org/10.3758/s13428-019-01254-w>

- 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>
- Connell, L., & Lynott, D. (2016). Embodied semantic effects in visual word recognition. In M. H. Fischer & Y. Coello (Eds.), *Foundations of embodied cognition: Conceptual and Interactive Embodiment* (1st ed., Vol. 2, pp. 71–92). Routledge. <https://doi.org/10.4324/9781315751962-10>
- 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 able to run any psycholinguistic experiments at all in 1990? *Cognition*, 10(1–3), 65–70. [https://doi.org/10.1016/0010-0277\(81\)90026-3](https://doi.org/10.1016/0010-0277(81)90026-3)
- 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>
- 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>
- 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., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (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., 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>
- Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (3rd ed.). Sage. <https://socialscienc.es.mcmaster.ca/jfox/Books/Companion/>
- 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>
- 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., 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>
- 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>
- 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>
- 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>
- 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>
- Hvitfeldt, E. (2021). *paletteer: Comprehensive Collection of Color Palettes*. version 1.3.0. <https://github.com/EmilHvitfeldt/paletteer>
- 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>
- Juhasz, B. J., Yap, M. J., Dicke, J., Taylor, S. C., & Gullick, M. M. (2011). Tangible Words are Recognized Faster: The Grounding of Meaning in Sensory and Perceptual Systems. *Quarterly Journal of Experimental Psychology*, 64(9), 1683–1691. <https://doi.org/10.1080/17470218.2011.605150>
- 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>
- 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(1), 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. M. (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. <https://doi.org/10.4074/s00035033070101030>
- 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>
- Locher, R. (2020). *IDPmisc: "Utilities of Institute of Data Analyses and Process Design (www.zhaw.ch/idp)." R package version 1.1.20.* <https://cran.r-project.org/package=IDPmisc>
- 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>
- 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>
- Miceli, A., Wauthia, E., Lefebvre, L., Ris, L., & Simoes Loureiro, I. (2021). Perceptual and Interoceptive Strength Norms for 270 French Words. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.667271>
- 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>
- Müller, K. (2020). *here: A Simpler Way to Find Your Files.* R package version 1.0.1. <https://cran.r-project.org/package=here>
- Muraki, E. J., Doyle, A., Protzner, A. B., & Pexman, P. M. (2023). Context matters: How do task demands modulate the recruitment of sensorimotor information during language processing? *Frontiers in Human Neuroscience*, 16. <https://doi.org/10.3389/fnhum.2022.976954>
- 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>
- 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>
- 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>
- 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>
- 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>
- Pexman, P. M. (2012). Meaning-based influences on visual word recognition. In *Visual word recognition: Meaning and context, individuals and development* (Vol. 2, pp. 24–43). Psychology Press.
- Pexman, P. M. (2020). How does meaning come to mind? Four broad principles of semantic processing. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, 74(4), 275–283. <https://doi.org/10.1037/cep0000235>
- 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>
- 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>
- R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.1.2). R Foundation for Statistical Computing.
- RStudio Team. (2022). *RStudio: Integrated Development Environment for R.* Version 2022.7.0.548.
- Siakaluk, P. D., Pexman, P. M., Aguilera, L., Owen, W. J., & Sears, C. R. (2008). 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. (2008). The Benefits of Sensorimotor Knowledge: Body-Object Interaction Facilitates Semantic Processing. *Cognitive Science*, 32(3), 591–605. <https://doi.org/10.1080/03640210802035399>
- Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arachchige, C., Arppe, A., Baddeley, A., Barton, K., Bolker, B., Borchers, H. W., Caeiro, F., Champely, S., Chessel, D., Chhay, L., Cooper, N., Cummins, C., Dewey, M., Doran, H. C., & Zeileis, A. (2022). *DescTools: Tools for descriptive statistics.* R package version 0.99.45. <https://cran.r-project.org/package=DescTools>

- Speed, L. J., & Majid, A. (2017). Dutch modality exclusivity norms: Simulating perceptual modality in space. *Behavior Research Methods*, 49(6), 2204–2218. <https://doi.org/10.3758/s13428-017-0852-3>
- 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>
- 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>
- Tastle, W. J., & Wierman, M. J. (2007). Consensus and dissent: 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>
- 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>
- 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>
- 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>
- 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>
- Vergallito, A., Petilli, M. A., & Marelli, M. (2020). Perceptual modality norms for 1,121 Italian words: A comparison with concreteness and imageability scores and an analysis of their impact in word processing tasks. *Behavior Research Methods*, 52(4), 1599–1616. <https://doi.org/10.3758/s13428-019-01337-8>
- 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>
- Wickham, H., & Bryan, J. (2023). *readxl: Read Excel Files. R package version 1.4.2*. <https://cran.r-project.org/package=readxl>
- Wickham, Hadley, Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.2110/5/joss.01686>
- Winter, B., Lupyan, G., Perry, L. K., Dingemanse, M., & Perlman, M. (2023). Iconicity ratings for 14,000+ English words. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-023-02112-6>
- 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>
- 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. J. (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>
- 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>

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/84564-addressing-the-elephant-in-the-middle-implications-of-the-midscale-disagreement-problem-through-the-lens-of-body-object-interaction-ratings/attachment/173637.docx?auth_token=g5-okREQ-R-Zhaf39aKL

Supplemental Material

Download: https://collabra.scholasticahq.com/article/84564-addressing-the-elephant-in-the-middle-implications-of-the-midscale-disagreement-problem-through-the-lens-of-body-object-interaction-ratings/attachment/173638.docx?auth_token=g5-okREQ-R-Zhaf39aKL
