Psychological Research (2018) 82:787–805
https://doi.org/10.1007/s00426-017-0864-8

ORIGINAL ARTICLE

A new statistical model for analyzing rating scale data pertaining to word meaning

Felipe Munoz-Rubke[1,2,3] · **Karen Kafadar**[4] · **Karin H. James**[1,2,5]

Received: 28 October 2016 / Accepted: 31 March 2017 / Published online: 25 April 2017

Abstract The concrete-abstract categorization scheme has guided several research programs. A popular way to classify words into one of these categories is to calculate a word's mean value in a Concreteness or Imageability rating scale. However, this procedure has several limitations. For instance, results can be highly distorted by outliers, ascribe differences among words when none may exist, and neglect rating trends in participants. We suggest using an alternative procedure to analyze rating scale data called median polish analysis (MPA). MPA is tolerant to outliers and accounts for information in multiple dimensions, including trends among participants. MPA performance can be readily evaluated using an effect size measure called analog R^2 and be integrated with bootstrap 95% confidence intervals, which can prevent assigning inexistent differences among words. To compare these analysis procedures, we asked 80 participants to rate a set of nouns and verbs using four different rating scales: Action, Concreteness, Imageability, and Multisensory. We analyzed the data using both two-way and three-way MPA models. We also calculated 95% CIs for the two-way models. Categorizing words with the Action scale revealed a continuum of word meaning for both nouns and verbs. The remaining scales produced dichotomous or stratified results for nouns, and continuous results for verbs. While the sample mean analysis generated continua irrespective of the rating scale, MPA differentiated among dichotomies and continua. We conclude that MPA allowed us to better classify words by discarding outliers, focusing on main trends, and considering the differences in rating criteria among participants.

Electronic supplementary material The online version of this article (doi:10.1007/s00426-017-0864-8) contains supplementary material, which is available to authorized users.

✉ Felipe Munoz-Rubke
lfmunoz@indiana.edu

[1] Cognitive Science Program, Indiana University Bloomington, 1101 E. 10th Street, Bloomington, IN 47405, USA

[2] Program in Neuroscience, Indiana University, Bloomington, USA

[3] Department of Statistics, Indiana University, Bloomington, USA

[4] Department of Statistics, University of Virginia, Charlottesville, USA

[5] Department of Psychological and Brain Sciences, Indiana University, Bloomington, USA

Introduction

We learn about the world through our bodily actions, as we act on objects, spaces, and other living beings. Throughout an uninterrupted interface with the environment, we collect sensorimotor data that shape our cognitive apparatus and its functioning (Varela, Thompson, & Rosch, 2000). Popular embodied theories maintain that cognitive processes are synonymous with sensorimotor interactions (Barsalou, 2008; Varela et al., 2000), such that there is no need of postulating the existence of cognitive processes operating independently of them.

Not all researchers agree on this, as some of them believe that our knowledge of the world is not grounded in sensorimotor data, but it is instead abstracted from direct experience. This process of abstraction results in the creation of amodal representations that result from direct experience, but become independent of said experience through internal interactions with other stored representations (Mahon & Caramazza, 2008). This latter perspective

has been popular among investigators in the field of word meaning (Collins & Loftus, 1975; Levelt, Roelofs, & Meyer, 1999), as words are symbols that denote our experience of the world, but are arbitrarily linked to their referents (Levelt et al., 1999).

However, the idea that words vary in the degree to which they represent concrete objects or actions vs. abstract concepts is an appropriate way to connect words back to the physical characteristics of their referents. The distinction between concrete and abstract words has been part of language theory for many years (Paivio, Yuille, & Madigan, 1968) and has guided numerous research programs (Allen & Hulme, 2006; Binder, Westbury, McKiernan, Possing, & Medler, 2005; Crutch, Troche, Reilly, & Ridgway, 2013; de Groot, 1989; Jessen et al., 2000; Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011; Kroll & Merves, 1986; Romani, Mcalpine, & Martin, 2008; Schwanenflugel, Akin, & Luh, 1992; Schwanenflugel & Shoben, 1983; Troche, Crutch, & Reilly, 2014). Probably the most remarkable achievement of this categorization scheme has been the 'concreteness effect', a cognitive advantage for concrete over abstract words in tasks involving word recognition and recall, word naming, sentence comprehension, among other measures (for review see Paivio, 1991). However, despite its popularity, the 'concreteness effect' has not been consistently reported in the literature (Fiebach & Friederici, 2004; Papagno, Fogliata, Catricalà, & Miniussi, 2009; Sabsevitz, Medler, Seidenberg, & Binder, 2005; Tsai et al., 2009; Tyler, Russell, Fadili, & Moss, 2001), thus raising some questions about the actual differences between concrete and abstract words.

The distinction between concrete and abstract words assumes different representational formats and/or contents for each word type. While the meaning of concrete terms relies on identifiable referents with clear distinguishable physical properties, abstract words are more detached from sensory experience and defined more by their connection to other words than their concrete counterparts (Borghi et al., 2017). But a strict boundary between concrete and abstract words may be an oversimplification, as perhaps a more graduated distribution of concrete to abstract reference could better capture the representation of word meaning. Prior to delving further into this possibility, we first outline the current work on classification of word meaning based on the concrete-abstract dichotomy.

Although there are several issues surrounding the classification of words into these categories, we address only three of these here. The first issue involves several theories on the constitution of concrete and abstract categories; the second relates to use of rating scales for placing words into these categories; the third concerns the statistical procedures used for such categorization.

Theories of concrete and abstract word meaning

There is some agreement that concrete and abstract words differ in terms of perceivability. While concrete words represent tangible entities, actions, or situations that can be perceived through the senses, the meaning of abstract words is variable, is not directly built upon sensory information and is partially understood by association to other words (Borghi et al., 2017; Wiemer-Hastings, Krug, & Xu, 2001).

The stored representations of words contain information that is used to encode and later recognize, identify, and use them. Depending on the theoretical viewpoint, representations of word meaning may be more or less reflective of how words are acquired and, therefore, may assign a role to sensorimotor information as part of a word′s representation (for review see Borghi et al., 2017).

At one extreme, latent semantic analysis (LSA) (Landauer, 1999; Landauer & Dumais, 1997) suggests that word meaning—at least most of it—is derived from our contact with words alone and is therefore independent of the sensorimotor information associated with a word's referent. LSA, which estimates semantic similarity by calculating distances among abstract symbols in a *k*-dimensional space, represents an amodal or 'unembodied' theory. LSA is only one of the statistical techniques used to infer meaning from linguistic corpora. Other renowned methods based on the same general logic, but not on the same procedures, are the Hyperspace Analogue to Language (HAL) (Lund & Burgess, 1996), Probabilistic LSA (Hofmann, 1999), the Topic model (Griffiths & Steyvers, 2004), and the BEAGLE model (Jones & Mewhort, 2007), among others. Proponents of text corpus analyses have suggested that while abstract words are associated to rich linguistic contexts, concrete words are connected to rich physical contexts describing objects and locations (Recchia & Jones, 2012).

The context availability theory (CAT) (Schwanenflugel et al., 1992; Schwanenflugel & LaCount, 1988; Schwanenflugel & Shoben, 1983) also suggests that the difference between concrete and abstract words occurs in terms of contexts. Instead of emphasizing a distinction between linguistic and physical contexts, CAT focuses on the difficulty of accessing world knowledge during word meaning processing. CAT claims that concrete words are strongly associated to a limited number of contexts whereas abstract words are loosely linked to a more diverse number of them. This would make abstract words difficult to retrieve from previous knowledge and would explain why they are, in general, more difficult to process.

Different from CAT, the dual coding theory (DCT) (Paivio, 1986, 1991) stresses the importance of sensory information in the distinction of concrete and abstract words. DCT suggests that both word types have verbal–symbolic representations, but that concrete words are also grounded in perceptual data. DCT claims that there is a positive correlation between concrete words and imageability, as concrete words can easily be linked to images (Borghi et al., 2017). However, recent work has criticized the study of perceptual experience through imageability measurements because they could inappropriately reduce the former to visual experience alone (Connell & Lynott, 2012).

At the other extreme, embodied theories propose that sensorimotor information plays a significant role in the meaning of both concrete and abstract words. For example, the language and situated simulation (LASS) theory (Barsalou, 2008) claims that meaning is based on experience and, therefore, grounded in modality-specific contexts. According to LASS, meanings are partially made up of information coming from different domains—such as perception, action, and internal states—that is used to simulate the state the brain would be in if each word's referent were to be experienced (Kiefer & Barsalou, 2013). By highlighting the role of simulation, this theory does not neglect the linguistic system's role in the construction of word meaning, but instead suggests that both linguistic and simulation components contribute to its conception. The words as social tools (WAT) theory (Borghi & Binkofski, 2014; Borghi, Scorolli, Caligiore, Baldassarre, & Tummolini, 2013) also recognizes the importance sensorimotor systems have in the generation of word meaning, at the same time that emphasizes that all linguistic experiences occur within social contexts. According to WAT, the meanings of both concrete and abstract words are simultaneously based on our social, linguistic, and physical interactions (Borghi et al., 2017).

Different approaches stress the importance of distinct components in the differentiation of concrete and abstract words. Although researchers mostly agree on perceivability as one of the factors distinguishing them (Wiemer-Hastings et al., 2001), the interpretation of what perceivability entails, or even whether perceivability is the only relevant factor, may change depending on the theoretical perspective.

Additionally, proponents of embodied perspectives emphasize that information beyond perceivability may also be relevant to word meaning, like the actions we use to interact with objects and spaces referred by words. Defining words in terms of how they are learned through physical experience could allow differentiation among them in situations where perceivability differences are not necessarily applicable (Barsalou, 2008; Borghi et al., 2013). For instance, when categorizing based on perceivability, we could assign no differences to words referring to objects used daily (e.g., pens) and words referring to objects we barely experience in the real world (e.g., volcano), because both words refer to entities with unmistakable, perceivable physical properties. In contrast, categorizing based on action could allow differentiating such words because people use pens, but they do not usually interact with volcanoes.

Regardless of the components that are relevant to the categorization of words, the procedure to classify them has been consistent among researchers. This procedure has relied on the use of rating scales that give participants instructions on how to rate words.

The use of rating scales for word categorization

A typical procedure for categorizing words uses rating scales completed by following specific instructions (Altarriba, Bauer, & Benvenuto, 1999; Binder et al., 2005; Fliessbach, Weis, Klaver, Elger, & Weber, 2006; Sabsevitz et al., 2005). Two popular instruments are the Concreteness and Imageability scales. The Concreteness scale measures the degree of concreteness or abstractness that a word has (Altarriba et al., 1999), while the Imageability scale measures how easy it is to create a mental image of a word's referent (Altarriba et al., 1999; Paivio, 1986, 1991).

Despite their widespread use, recent evidence suggests that neither scale reflects the perceptual experience connected to word processing. While in the case of the Concreteness scale it is unclear what type of information is used to guide the classification, the Imageability scale disregards the complexity of perceptual experiences (e.g., personal experience vs. seeing images in books) and focuses almost exclusively on available visual information (Connell & Lynott, 2012).

Alternatively, other work has highlighted the relevance of sensorimotor information for word classification. For instance, Tillotson, Siakaluk, & Pexman (2008) created body–object interaction (BOI) ratings by asking participants to evaluate how easy it was for the human body to interact with entities referred by nouns. An extension of this approach was conducted by Sidhu, Kwan, Pexman, & Siakaluk (2014), who asked participants to rate verbs based on how often an action, state, or relation involved the human body. In the same vein, Amsel, Urbach, & Kutas (2012) asked participants to rate nouns with respect to the sensory (e.g., sound, taste) and motor (e.g., graspability) attributes of their respective referents.

These instruments represent only a small sample of the scales used by researchers. Other scales focus on different dimensions such as emotion, valence, arousal, polarity, morality, and motion, among others (Hoffman & Lambon Ralph, 2013; Kousta et al., 2011; Troche et al., 2014). By using rating scales, investigators have measured diverse aspects of words and, depending on the type of statistical analysis used, these scores have been used to emphasize more continuous or more dichotomous classifications of word meaning.

Statistical procedures

Although words often are ascribed as being either concrete or abstract, researchers acknowledge that a concrete–abstract dichotomy is only partially correct (Borghi, Flumini, Cimatti, Marocco, & Scorolli, 2011; Ghio, Vaghi, & Tettamanti, 2013). Previous studies have shown that concreteness ratings obtained for a large sample of nouns form a continuum of meaning characterized by a bimodal distribution (Nelson & Schreiber, 1992; Wiemer-Hastings et al., 2001). However, despite the theoretical acceptance of this continuum, some studies have treated concrete and abstract terms in a dichotomous fashion (e.g., Binder et al., 2005; Jessen et al., 2000; Papagno, Martello, & Mattavelli, 2013).

This mismatch between theory and practice may be fostered by the widespread use of the Concreteness and Imageability scales, which by disregarding the complexity of perceptual experiences could result in oversimplifying the complexity of word meaning (Connell & Lynott, 2012). This disparity could also be attributed to the statistical procedure used to analyze rating scale data, in which it has been common practice to estimate a distribution of word meaning by: (1) asking participants to rate words based on some specific criteria (e.g., how easy it is to evoke a mental image of the entity or action referred by a word?); (2) calculating the mean rating value for each word; (3) creating histograms that reveal the distribution of these mean rating values; (4) deciding on a cutoff value to distinguish between concrete and abstract words (Nelson & Schreiber, 1992; Wiemer-Hastings et al., 2001).

However, following these steps could lead to misleading findings for several reasons. First, this procedure disregards the controversy of using means to describe ordinal variables, which is important in the case of rating scale data (Jamieson, 2004; Sullivan & Artino, 2013). A second issue relates to the lack of robustness of the sample mean, which can be highly distorted by atypically small or large values. In these situations, the mean could suffer from inflation or deflation due to the influence of extreme outliers. Third, by promoting the use of point estimates (e.g., means) instead of interval estimates (e.g., confidence intervals), this procedure could artificially lead to ascribing differences among observations when none exists. Fourth, this procedure ignores the variability among participants. Participants can rate words with atypical scores due to different response set biases, such as lack of interest and misunderstanding of rating scale norms, or due to idiosyncratic perspectives on word meaning. Regardless of the motives driving the behaviors of participants, ignoring the variability among them can only have negative effects.

Current project

The current work focuses on the statistical analysis of rating scale data and offers an alternative to the aforementioned procedure based on mean rating values. The proposed method uses a robust statistical technique called median polish analysis (MPA) (Hoaglin, Mosteller, & Tukey, 2000, 2006; Tukey, 1977) that introduces several benefits when compared to more traditional approaches. First, it does not rely on means, but on iteratively finding and subtracting row and column medians along relevant dimensions of data sets. Second, MPA is more tolerant to the presence of outliers than both the mean and other linear methods based on it (e.g., *t* test, ANOVA, etc.) and does not necessitate the removal of outliers by hand. This avoids introducing biases through the subjective and arbitrary elimination of data points. Third, MPA can be easily integrated with the creation of bootstrap confidence intervals (CIs). Using CIs allows a transition from point to interval estimations that lends valuable information for replication (Cumming, 2008) and reduces the temptation to ascribe nonexistent differences among words. Fourth, MPA performance can be evaluated via an effect size measure called analog R^2. The importance of using effect sizes has been emphasized in the statistical literature, as its use could significantly improve the quality of research (Cumming, 2014). Fifth, MPA can account for both individual word ratings and individual participant preferences because it analyzes data distributed across both dimensions. It does this by generating scores (or effects) for each word and each participant.

Here, we introduce MPA by presenting some of the information it can generate after exploring data sets organized along 2 or 3 dimensions. For instance, we show that two-way MPA models can explore differences among words by assigning effects to each one of them. In doing so, MPA results suggest whether certain rating scales promote the creation of continuous or dichotomous distributions of meaning. We also show that three-way MPA

models can explore trends present within both word categories and participants, and by doing this they can indicate whether different rating scales produce dissimilar effects along these dimensions.

Consistent with the importance of using CIs and effect size measures for improving the quality of research (Cumming, 2014), we also demonstrate how MPA can be integrated with the creation of both bootstrap CIs and the calculation of an analog R^2. Because the main concern in studies of word meaning relates to the position of each word within a given rating scale, we show how bootstrap CIs can be estimated for each word.

We used classical multidimensional scaling (CMDS) (Everitt & Hothorn, 2011), a well-known technique, to validate the results of MPA presented here. Although it is expected for both MPA and CMDS to generate similar outcomes in data sets that do not contain gross outliers, results obtained with linear techniques like CMDS, Principal Component Analysis (PCA) or Factor Analysis (FA) can be non-robust and, therefore, dramatically distorted by the presence of even a single extreme outlier (Candès, Li, Ma, & Wright, 2011; Spence & Lewandowsky, 1989). Because it is not possible to know in advance whether or not data sets will contain gross outliers, the use of robust procedures like MPA should always be encouraged. This is particularly relevant for word meaning data, as such data can include large numbers of participants and words. An additional benefit of MPA over CMDS is that the former keeps the results in the same scale as the original data, while CMDS transforms these values into new dimensions that researchers must later interpret.

In line with our main goal of showing how MPA can be used to analyze rating scale data, as well as presenting an example of the output it offers to researchers, we asked 80 participants to rate the same nouns and verbs using four different rating scales: Concreteness, Imageability, Multisensory, and Action. We then handled and processed these data using both a standard procedure (sample mean), an MPA procedure, and a CMDS procedure. In this context, we were particularly interested in knowing the circumstances under which continuous distributions of meaning could be identified and in determining the efficacy of the different statistical models in describing the data.

We selected the Concreteness and Imageability scales due to their widespread use and popularity. We also chose the Imageability scale because it has been claimed that the main difference between concrete and abstract words occurs in terms of imageability (Paivio, 1986). The Multisensory scale was added to capture the complexity of perceptual information and avoid its reduction to visual images, as it could happen with the Imageability scale (Connell & Lynott, 2012). The Multisensory scale resembles the work of Amsel et al. (2012) and Connell & Lynott (2012), but instead of asking separate questions for each sensory modality considers them all simultaneously. We also included an Action scale to account for the role of sensorimotor information in word meaning processing. As advanced by embodied accounts of word meaning (Barsalou, 2008; Borghi et al., 2013), this information could be useful in distinguishing more concrete from more abstract words. The Action scale used here resembles the BOI ratings previously utilized in the literature (Sidhu et al., 2014; Tillotson et al., 2008). Our selection of rating scales was not exhaustive and it was not intended to represent all dimensions relevant to word meaning processing. We selected these rating scales because they were popular in the literature and linked to dimensions that have been highlighted by theories of concrete and abstract words. In other words, the procedures described here could easily be extended to any other rating scales of interest.

The words chosen to demonstrate the use of MPA were selected from four different categories of nouns and verbs. Because the selection of these categories was guided for the purpose of demonstrating how to use MPA, it was not exhaustive, and it did not represent all possible categories of either nouns or verbs. Instead, our choice of words was guided by the need to generate clear distinctions among categories.

In the case of nouns, we used four categories (for more details see the "Methods"): Manipulable (M), Non-Manipulable (NM), Social Organization (SO), and Abstract (A). The M, NM, and SO categories represented nouns whose referents were easily observable, while the A category comprised nouns that were more difficult to link to single, identifiable referents. The first three categories also differed from each other, with M words being more manipulable than both NM and SO words, and by NM words representing distinguishable physical entities and SO words representing culturally relevant distinctions or creations. We followed a similar procedure for verbs. The four verb categories selected were: Human Action (HA), Non-Human Action (NHA), Emotion (E), and Cognition (C). The HA, NHA, and E categories represented verbs whose referents were easily identifiable from the perspective of an external observer, while the C category comprised verbs that could be more difficult to distinguish from the perspective of an external observer. HA verbs referred to simple actions, commonly performed by humans; NHA verbs represented actions that humans could emulate, but are not part of their daily repertoire; E verbs represented emotional states; and C verbs represented cognitive states or mental activities (Rodríguez-Ferreiro, Gennari, Davies, & Cuetos, 2011).

We distinguished between nouns and verbs during word selection because they represent different word classes

(i.e., while nouns represent objects and entities, verbs represent actions and states), but this distinction need not be made when analyzing data with MPA. Indeed, the three-way MPA models presented here analyzed all categories together, irrespective of whether they corresponded to noun or verb categories; whereas the two-way MPAs presented here analyzed nouns and verbs separately from each other. There is no principled reason why nouns and verbs could not be analyzed together using any variant of MPA. This is a decision left to the researcher, based on his/her specific research questions and hypotheses.

Although MPA is not a novel technique and it is certainly familiar among data analysts (Hoaglin et al., 2000, 2006; Tukey, 1977), it has not been applied in the field of word meaning before. To promote a better understanding of this technique and to encourage its use in future studies, we have included an R tutorial in the Additional Materials that we hope will help researchers analyze and interpret their own data with MPA.

Methods

Participants

Eighty undergraduate students (mean age 23.5 years, age range 22–30 years, 60 females) at Indiana University Bloomington took part in the experiment for course credit. All of them were native English speakers, and had normal or corrected-to-normal visual acuity. Informed consent was obtained from each participant before the experiment, in accordance with the IUB Institutional Review Board approved protocol.

Materials

Stimuli

A total of 136 words (68 nouns and 68 verbs) were used for the purposes of the current experiment. Nouns were selected from the following categories: Manipulable (M), Non-Manipulable (NM), Social Organization (SO), and Abstract (A). The M category encompassed easily observable and manipulable objects (e.g., ball, pencil), while the NM category comprised easily observable but not easily manipulable objects (e.g., palace, moon). The SO category incorporated nouns related to social organization and culturally relevant creations, such as occupations and seasons of the year. SO nouns could easily be connected to visual representations that physically resembled their respective referents (e.g., doctor, winter). Finally, the A category contained nouns that could not be easily connected to a visual representation that physically resembled the word's referent (e.g., origin, regret). Each category had 17 words (68 nouns total).

Verbs were selected from the following categories: Human Action (HA), Non-Human Action (NHA), Emotion (E), and Cognition (C). The HA category encompassed actions commonly performed by humans in their everyday lives (e.g., to comb, to hug), while the NHA category comprised actions that can be emulated by humans, but are mostly part of the everyday lives of non-human animals (e.g., to bark, to moo). The E category incorporated verbs describing emotions (e.g., to love, to panic), while the C category contained verbs referring to a variety of mental processes different from emotions (e.g., to confuse, to think). Each category had 17 words (68 verbs total).

Word length and frequency

Information on word length and frequency of use was obtained from the English Lexicon Project (Balota et al., 2007). A one-way MANOVA was conducted to assess differences among noun categories in terms of word length and frequency of use. Results showed no statistically significant differences among these groups [$F(3,64) = 0.41, p > 0.05$]. Another one-way MANOVA was conducted to test differences among verb categories in terms of word length and frequency of use. In this case, a statistically significant MANOVA effect was obtained, [$F(3,64) = 11.46$, $p < 0.01$]. Bonferroni-corrected post hoc tests showed that the Non-Human Action (NHA) category differed from the remaining categories in terms of frequency of use [vs. HA: $t(32) = 3.85, p < 0.01$; vs. E: $t(32) = 3.28, p < 0.05$; vs. C: $t(32) = 3.87$, $p < 0.01$]. This result was expected because NHA words are usually infrequent in everyday conversations (e.g., people do not talk often about barking or mooing). Bonferroni-corrected post hoc tests also showed differences among verb categories in terms of word length [C vs. HA: $t(32) = 5.72, p < 0.01$; C vs. NHA: $t(32) = 6.25, p < 0.01$; E vs. HA: $t(32) = 3.22, p < 0.05$; E vs. NHA: $t(32) = 3.69$, $p < 0.01$]. The only comparisons that did not show statistically significant differences in word length were those contrasting categories C and E [$t(32) = 2.5$, $p > 0.05$], and categories HA and NHA [$t(32) = 0.34$, $p > 0.05$]. This result was also expected because words belonging to HA and NHA categories were usually short in length [e.g., run (HA), bark (NHA)].

Because we did not measure response times (RTs) and there were no predefined correct answers for our task, our only concern was that participants would understand the words selected by us. With respect to this, word length differences were not considered to be an important factor. However, frequency of use may have affected our results (though, notice that differences in frequency of use only affected the NHA verbs). Nonetheless, we decided to keep

the NHA words for two reasons: first, because these words could provide valuable information concerning the action component of words referring to activities that we can emulate, but are not part of our typical behavioral repertoire; second, because we considered NHA words to be extremely simple to understand, as they are usually learned early in life.

Scale selection

Four rating scales were used in the present study: the two well-known Concreteness and Imageability scales, as well as the two newly created Multisensory and Action scales (see Table 1). Both the Concreteness and Imageability scales have been widely used to classify words in previous studies (Altarriba et al., 1999; Brysbaert, Warriner, & Kuperman, 2014; Fliessbach et al., 2006; Giesbrecht, Camblin, & Swaab, 2004; Sabsevitz et al., 2005).

We followed Altarriba's operationalization (Altarriba et al., 1999) and set the Concreteness scale to directly ask how concrete or abstract a word was, thus appealing to an individual's intuitive knowledge. On this scale, (1) represents a word that is very abstract, whereas (7) represents a word that is very concrete. The Imageability scale was also set based on Altarriba's operationalization (Altarriba et al., 1999) and asked how difficult or easy it was to evoke a mental image of a word's referent. On this scale, (1) represents a word that would be extremely difficult to imagine, whereas (7) represents a word that would be extremely easy to imagine. Because of the general nature of the directions provided by both scales, the instructions were identical for both nouns and verbs.

Both the Multisensory and Action scales were created for the purposes of the present experiment, but have strong connections with those used in previous studies (Amsel et al., 2012; Sidhu et al., 2014; Tillotson et al., 2008).

The Multisensory scale was conceived as an extension of the Imageability scale, as the latter predominantly targets the visual modality (Connell & Lynott, 2012). The purpose of the Multisensory scale was to emphasize information coming from all sensory modalities. On this scale, (1) represented a word that would be extremely difficult to relate to a sensory experience, whereas (7) represented a word that would be extremely easy to relate to a sensory experience.

The Action scale asked participants to rate words based on how easy or difficult it was to manipulate a noun's referent with their hands (nouns) and how easy or difficult it was to perform the activity described by a verb (verbs). The decision to create two versions of the scale was connected to the difference between nouns and verbs, with nouns referring mostly to entities and verbs referring mostly to actions and states. For nouns, (1) represented a noun that would be extremely difficult to manipulate with the hands, whereas (7) represented a noun that would be extremely easy to manipulate with the hands. For verbs, (1) represented an activity that would be extremely difficult to perform, whereas (7) represented an activity that would be extremely easy to perform.

Table 1 Instructions given in each rating scale and examples provided to participants

Instructions given to participants			
Scales	Instructions	N/V	EXAMPLE given to participants
Action	Can you manipulate it with your hands?	Nouns	TRUTH = rated as 1 or 2 SCREWDRIVER = rated as 6 or 7
	Is it an activity you can perform?	Verbs	To TRUST = rated as 1 or 2 To SWIM = rated as 6 or 7
Concreteness	How concrete or abstract do you think it is?	Nouns	TRUTH = rated as 1 or 2 SCREWDRIVER = rated as 6 or 7
		Verbs	To TRUST = rated as 1 or 2 To SWIM = rated as 6 or 7
Imageability	Can you form a mental image of it?	Nouns	TRUTH = rated as 1 or 2 SCREWDRIVER = rated as 6 or 7
		Verbs	To TRUST = rated as 1 or 2 To SWIM = rated as 6 or 7
Multisensory	Can you relate it to a sensory experience (taste, touch, sight, sound, or smell)?	Nouns	TRUTH = rated as 1 or 2 SCREWDRIVER = rated as 6 or 7
		Verbs	To TRUST = rated as 1 or 2 To SWIM = rated as 6 or 7

Testing stimuli

During testing, each word was displayed in Arial font (approximately size 70) and centered on a Dell™ Professional™ P1911 48.26 cm (19″) W monitor. The words were presented using Psychopy (Pierce, 2007) and subtended 4.7 degrees of visual angle on average.

Design

The four scales and the two word classes resulted in eight different experimental blocks: Action/Noun, Action/Verb, Concreteness/Noun, Concreteness/Verb, Imageability/Noun, Imageability/Verb, Multisensory/Noun, and Multisensory/Verb. Trials were blocked to simplify the task for participants. Within each block, trials were randomized. Presentation order of blocks was pseudo-randomized among participants to alleviate possible order effects.

Procedure

Each testing block began with instructions that included the criteria to rate words, a description of what would mean to assign extreme values to a word, and a simple example (see Table 1). The instructions/examples were given at the beginning of each block and before any stimuli were presented. After reading the instructions, participants pressed a key to confirm they understood them. Participants were also told that nouns and verbs would be separated into different blocks and that they would observe a new set of instructions before starting a new block. To avoid confusion, nouns were presented in their singular form (e.g., screwdriver) and verbs were presented in their infinitive form (e.g., to swim). Participants responded by keypress and were told to respond as accurately as possible. Additionally, they were told that they could take as much time as needed.

The first trial was presented 500 ms after the instructions and examples. The word would appear on the computer screen until a rating response was given. Participants responded by pressing the keys 1 through 7 on a standard QWERTY keyboard. Following the response, a 500-ms break occurred between trials. All 68 nouns or verbs were included in each block.

Median polish analysis (MPA)

We analyzed the data using different techniques: three-way MPA, two-way MPA, CMDS, and sample mean. We also created 95% bootstrap CIs for each word based on two-way MPA models.

Median polish analysis (MPA) is a robust exploratory technique created by John Tukey, particularly suited for the analysis of 2D and 3D matrices (Hoaglin et al., 2000, 2006; Tukey, 1977). It is a robust method because of its high resistance to outliers, and exploratory because it emphasizes a flexible investigation of evidence. Instead of prioritizing an evaluation of the evidence in terms of its extrapolation to the entire population, as it is the purpose of a confirmatory analysis, exploratory data analysis techniques try to find patterns in the data without being influenced by exotic values or outliers.

MPA is a simple additive model that summarizes the connection between a dependent variable and the joint contribution of factors (Hoaglin et al., 2000, 2006; Tukey, 1977). In its 2D version, this additive model takes the following form (Eq. 1):

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \quad (1)$$

y_{ij} represents each of the entries in a 2D matrix (dependent variable); μ is the overall value for the entire model or 'grand effect' and represents the median value across all dimensions; α_i represents the 'row effect' or the contribution of the variable described by row; β_j represents the 'column effect' or the contribution of the variable described by column; γ_{ij} is the error term. Because the model is additive, y_{ij} can be reconstructed by adding the grand effect together with the corresponding row effect, column effect, and error term.

The robustness of MPA is reflected in its breakdown value. Breakdown values describe the smallest number of outliers that can be present in a data set before having a severely distorting effect over an estimator (Hubert & Debruyne, 2009), and as such reflect the robustness of statistical procedures. While the breakdown value of the sample mean is 0, given that the presence of a single outlier can distort it and affect its interpretability, the breakdown value of the MPA model is approximately min (I/2, J/2). Here, 'I' represents the number of row entries and 'J' represents the number of column entries (Hoaglin et al., 2000, 2006). Put simply, MPA is more robust to outliers than the sample mean, but cannot deliver reliable information if the number of outliers in the data set exceeds its breakdown value.

The MPA procedure consists of iteratively finding and subtracting row medians and column medians from the data set, until all rows and columns have a zero median. This technique creates row effects (α_i) for the variable organized along the x-axis and column effects (β_j) for the variable arranged along the y-axis. For instance, in the case of a 2D matrix where participants are organized by row (I = participants) and words are organized by column (J = words), MPA provides with an estimated effect for each participant and each word. For each axis, results take the form of one vector with the same number of elements as the number of participants (vector containing the α_i) and

another vector with the same number of elements as the number of words (vector containing the β_j).

The goodness of fit for the MPA model can be evaluated with an analog R^2. The analog R^2 receives its denomination for its similarity to the well-known R^2 statistic, and it also ranges from a minimum of 0 to a maximum of 1. The analog R^2 is calculated by the following formula:

$$\text{Analog } R^2 = 1 - \sum|\text{Residuals}| / \sum|\text{Original data} - \mu|. \quad (2)$$

Three-way MPA

For the purposes of the current project, we used four three-way MPA models to calculate effects along three dimensions for each rating scale: Participants (80 elements), Words (17 elements nested within each category), and Categories (8 elements: Noun/M, Noun/NM, Noun/SO, Noun/A, Verb/HA, Verb/NHA, Verb/E, Verb/C). However, out of these dimensions, we only focused on those referring to Participants and Categories. We did not analyze the dimension containing the Words because the three-way MPA models estimated their corresponding effects under the assumption that the same 17 words were nested within each of the 8 categories, a condition that was not satisfied by our data.

With the intention of getting rid of the Word effects as much as we could (i.e., obtaining Word effects from the three-way MPA models whose values were near 0), we conducted the following procedure: (a) we fitted two-way MPA models to each data set (eight data sets in total, one for each experimental block: Action/Noun, Action/Verb, Concreteness/Noun, Concreteness/Verb, Imageability/Noun, Imageability/Verb, Multisensory/Noun, and Multisensory/Verb); (b) we subtracted the Word effects obtained from a two-way MPA model from the respective original data set, which generated residual data sets that were devoid of Word effects; (c) we organized the residual data sets by rating scale and created 3D matrices that were connected to the *n*th rating scale (4 in total), where the *ijk*th entry corresponded to the residual rating provided by the *i*th Participant to the *j*th Word present within the *k*th category; (d) we fitted a three-way MPA model to each 3D matrix.

The results of each three-way MPA model were extracted in the form of three vectors. For each rating scale, the first vector contained 80 entries or effects (Participant effects), the second vector contained 17 entries or effects (Word effects), and the third vector contained 8 entries or effects (Category effects). These effects represented the existence of trends within each dimension (Hoaglin et al., 2000, 2006; Tukey, 1977). For example, if some participants were to strongly deviate from the grand effect (overall value) by rating words with excessively high or excessively low scores, we would expect for these trends to be captured by their respective Participant effects.

Subsequently, we explored whether Participant or Category effects differed across rating scales. We first grouped together the Participant effects obtained across all three-way MPA models, and we then did the same with the Category effects. To evaluate whether participants, as a group, assigned higher or lower values to some scales more than to others, we analyzed the Participant effects with a one-way ANOVA and with a Kruskal–Wallis test, a non-parametric alternative to the one-way ANOVA (Kloke & McKean, 2015). Both models included Scale (Action, Concreteness, Imageability, Multisensory) as their only predictor.

To evaluate whether word categories received higher or lower values depending on the rating scale, we analyzed the Category effects with both a two-way ANOVA and a rank-based two-way ANOVA, a robust implementation of ANOVA (Kloke & McKean, 2015). Both models included Scale and Word Class (Noun/Verb) as their only predictors. Although the three-way MPA models provided information about Participant and Category effects for each rating scale—information that we used to compare across scales with ANOVAs—the three-way MPA models did not give us information concerning each word. Namely, they did not tell us whether dichotomous or continuous distributions of word meaning could be identified by the different rating scales.

Two-way MPA

With the intention of obtaining effects for each word, we used eight two-way MPAs to calculate effects along two dimensions: Participants (80 elements) and Words (68 elements). This procedure was run separately for each experimental block: Action/Noun, Action/Verb, Concreteness/Noun, Concreteness/Verb, Imageability/Noun, Imageability/Verb, Multisensory/Noun, and Multisensory/Verb. The data were organized into several 2D matrices connected to the *n*th experimental block, where the *ij*th entry corresponded to the rating provided by the *i*th Participant to the *j*th Word. Results for each two-way MPA model were extracted in the form of two vectors, each representing a relevant dimension; that is, the first vector contained 80 entries or effects (Participant effects) and the second vector contained 68 entries or effects (Word effects). Because both Participant and Word effects were difficult to interpret directly, the grand effect (overall value) was added back to both effects via simple addition. Such procedure allowed us to resize the scores to a range between 1 and 7, which substantially facilitated interpretation.

Following this initial estimation, we then calculated 95% bootstrap CIs for each word. To do this, resampling with replacement was first used to artificially generate 1000 data sets. For each generated data set, a two-way MPA model was fitted and the corresponding effects were estimated. From the distribution of 1000 effects so calculated, we created the CIs by selecting the values corresponding to the 2.5 and 97.5% of this distribution.

To validate MPA, we used classical multidimensional scaling (CMDS). CMDS reduces dimensionality and represents proximities among words into a geometrical map (Everitt & Hothorn, 2011). We used CMDS to analyze the data sets associated with each experimental block (like we did with the two-way MPA models). We used both the trace and the magnitude criteria to determine if the reduction into two dimensions provided an adequate fit to the original data. While based on the trace criterion, the reduction into two dimensions was considered appropriate when the sum of the first two positive eigenvalues approximated the sum all eigenvalues; based on the magnitude criterion, the reduction into two dimensions was considered appropriate when the first two eigenvalues substantially exceeded the value of the largest negative eigenvalue (Everitt & Hothorn, 2011).

Finally, we calculated the mean value for each word by averaging across participants. We then sorted the values and plotted them against the MPA results (see Fig. 5).

All procedures were implemented using the R Language (2016). The procedures were run using in-house scripts that made use of 4 R packages: ggplot2 (Wickham, 2009), MASS (Venables & Ripley, 2010), Rfit (Kloke & McKean, 2015), and Rmisc. The R tutorials included in the Additional Materials replicate the same steps followed here for the implementation of both three-way and two-way MPA models.

Results

Three-way MPA

For the Action scale, the three-way MPA overall score was 4.5 and the analog R^2 for the model was 0.69. For the Concreteness scale, the overall score was 6 and the analog R^2 for the model was 0.66. For the Imageability scale, the overall score was 6 and the analog R^2 for the model was 0.7. For the Multisensory scale, the overall score was 4.9 and the analog R^2 for the model was 0.67. These overall values suggest that participants gave higher ratings to words with the Concreteness and Imageability scales than with the Action and Multisensory scales. They also show that all three-way MPA models were associated with strong effect sizes (all analog $R^2 > 0.5$). This means that the MPA models were good representations of the trends present within the original data sets.

To further explore differences among the rating scales, we evaluated whether the Participant effects differed among them. This provided information on whether participants, as a group, provided higher or lower ratings to words depending on the scale they used. To do this, Participant effects estimated via all three-way MPA models were grouped together and analyzed with a one-way ANOVA with Scale (Action, Concreteness, Imageability, and Multisensory) as the only predictor. A similar procedure was implemented using a Kruskal–Wallis test. Results from both statistical techniques showed a main effect of Scale [$F(3316) = 3.76$, $p = 0.011$; $H(3) = 10.17$, $p = 0.017$]. Post hoc comparisons using Tukey's HSD test suggested that participants gave lower values to words when using the Action scale than when using both the Concreteness scale ($p = 0.013$) and the Multisensory scale ($p = 0.034$). No other statistically significant differences between pairs of scales were found. Taken together, these results indicate that participants used the Action scale in a different way than they used the Concreteness and Multisensory scales, but they do not tell us much about how the remaining scales differ from one another (see Fig. 1).

We then evaluated differences among word categories across rating scales. Category effects were grouped together and analyzed with a two-way ANOVA with Scale (Action, Concreteness, Imageability, and Multisensory) and Word Class (Noun, Verb) as predictors. A similar procedure was implemented using a rank-based two-way ANOVA, a more robust implementation of ANOVA. Both procedures suggested a main effect of Category [$F(1,24) = 146.98$, $p < 0.001$; robust $F(1,24) = 139.21$, $p < 0.001$] and an interaction between Scale and Category [$F(3,24) = 37.32$, $p < 0.001$; robust $F(3,24) = 39.07$, $p < 0.001$]. No main effect of Scale was found [$F(3,24) = 0.57$, $p = 0.657$; robust $F(3,24) = -0.61$, $p \approx 1$].

Post hoc comparisons using Tukey's HSD test revealed that nouns were, in general, rated higher than verbs ($p < 0.001$). Both the Concreteness and Imageability scales showed identical results. For both scales, nouns were rated higher than verbs (both $p < 0.001$). Something similar was observed with respect to the Multisensory scale, as nouns were also rated higher than verbs with this instrument ($p < 0.001$). In contrast, the Action scale did not show a statistically significant difference between nouns and verbs ($p = 0.12$) (see Fig. 2).

Two-way MPA and CMDS

Nouns

The three-way MPA models provided information about participants and categories, but they did not tell us much

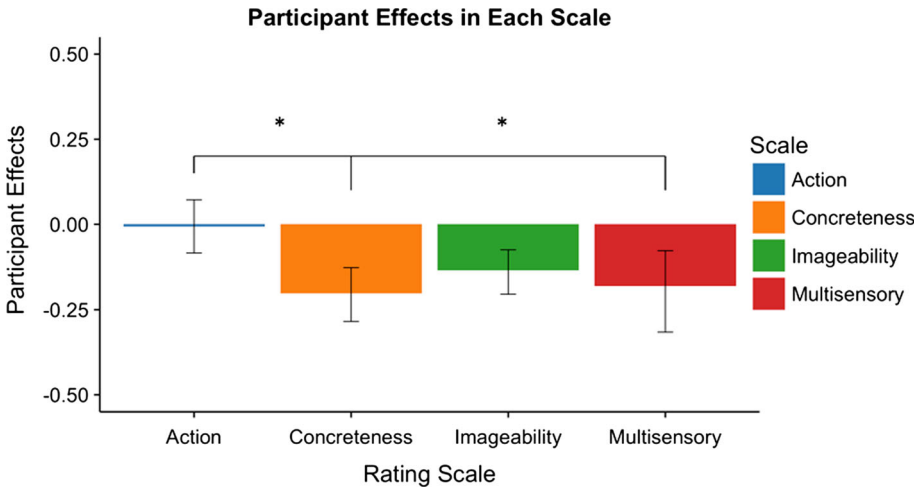

Fig. 1 *Bar plots* showing the estimated Participant effects. *Bars* depict the mean Participant effect value for each rating scale. *Error bars* represent 95% bootstrap CIs. Results indicate differences between the Action scale and both the Concreteness ($p < 0.05$) and Multisensory scales ($p < 0.05$)

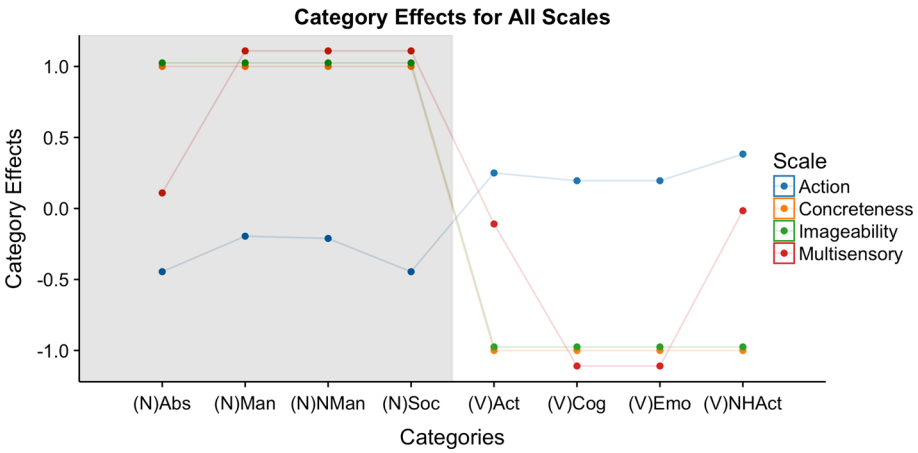

Fig. 2 Category effects estimated via three-way MPA. There were eight different Categories: Noun/Manipulable, Noun/Non-Manipulable, Noun/Social Organization, Noun/Abstract, Verb/Human Action, Verb/Non-Human Action, Verb/Emotion, and Verb/Cognition. In the figure, (N) stands for Noun (*shaded portion*) and (V) stands for Verb. Results for both the Concreteness and Imageability scales completely overlap (*orange* and *green lines*). When using the Concreteness ($p < 0.001$), Imageability ($p < 0.001$), and Multisensory scales ($p < 0.001$), participants assigned higher scores to nouns than to verbs. In contrast, the Action scale did not show a statistically significant difference between nouns and verbs ($p = 0.12$). These results, together with those obtained analyzing Participant effects, suggest that the Action scale focuses on a different underlying property of words than the remaining scales

about each word. To learn more about the distribution of single words that resulted from using each rating scale, we used two-way MPA models.

We considered a distribution as continuous whenever the words spanned most of the space located between extremely low and extremely high values. In contrast, we identified a distribution as dichotomous whenever the words tended to populate the extremes, or as stratified whenever the words were organized into several layers, each including at least ten elements.

In the case of the Action scale, results suggested the presence of a continuous classification. This could be observed in the results obtained by all three procedures (two-way MPA, 95% bootstrap CIs, and CMDS). The overall score for the two-way MPA was 4.2 and the analog R^2 for this model was 0.57, which represents a strong effect size. The plots for both the two-way MPA and the bootstrap CIs show that the words spanned the entire spectrum. These results agree with those found using CMDS (see Fig. 3, upper row). Both criteria used to evaluate fit of the CMDS model revealed that the first two dimensions provided an adequate representation of the data (trace = 0.784; magnitude = 0.994). In the first dimension, the Action scale created two poles. They were mainly populated by words from the Abstract and Manipulable categories. The remaining words were distributed between both extremes, suggesting the creation of a continuum. In turn, the second dimension seemed to differentiate words in the extremes of the first dimension from words in the middle of the distribution.

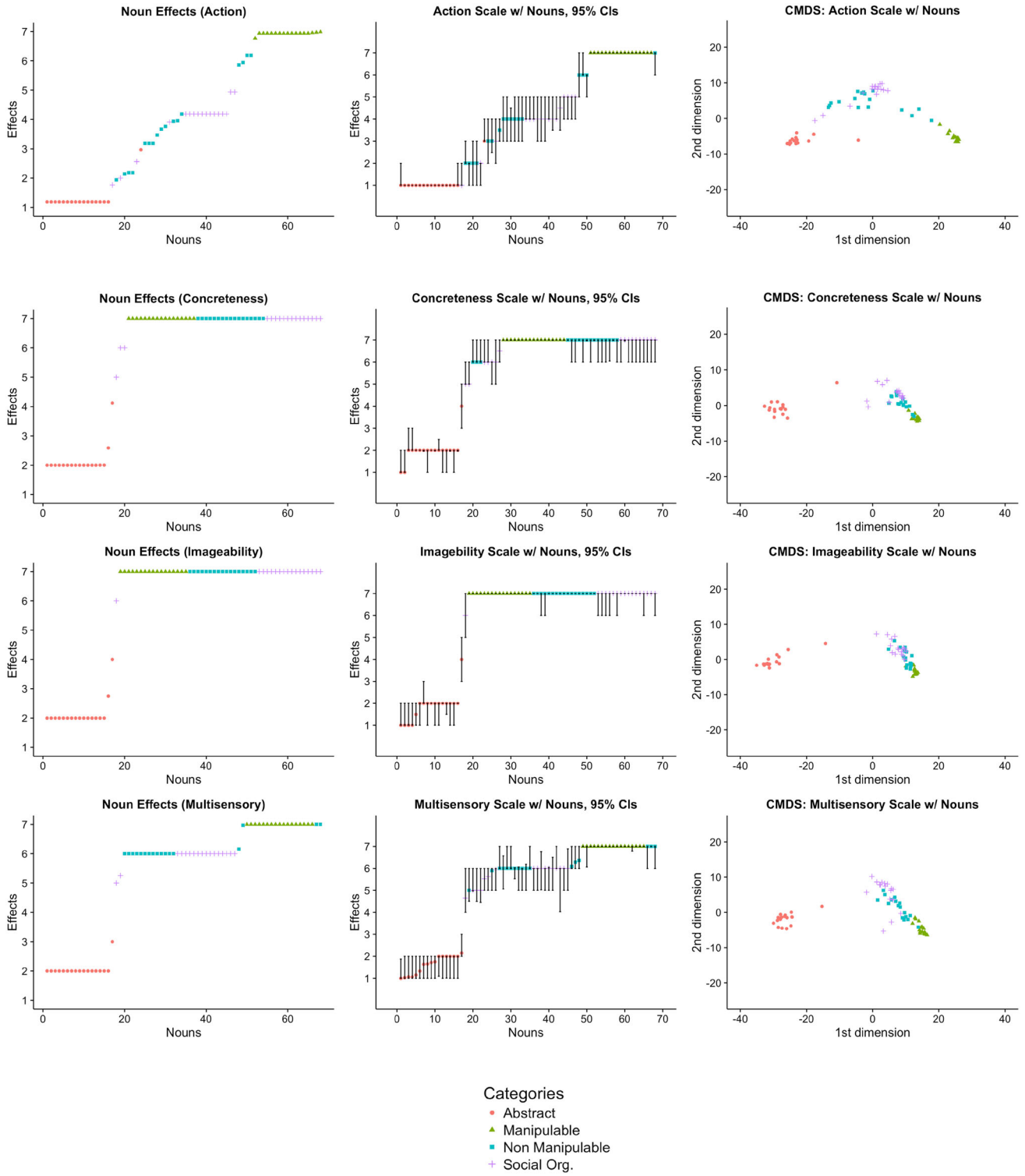

Fig. 3 Results from the two-way MPA models (*left column*), 95% bootstrap CIs (*middle column*), and CMDS models (*right column*) with Nouns. The *upper row* shows the results for the Action scale; the *middle upper row* shows the results for the Concreteness scale; the *lower middle row* shows the results for the Imageability scale; the *lower row* shows the results for the Multisensory scale. Each word has been colored and shaped according to its preassigned Category

In contrast, for both the Concreteness and Imageability scales, results suggested the presence of a dichotomous classification. This could be observed in the results obtained by all three procedures. Both MPA models had an overall score of 7. Additionally, both statistical models showed strong effect sizes, with an analog R^2 of 0.59 for the Concreteness scale data and an analog R^2 of 0.66 for the Imageability scale data.

For the Concreteness scale, plots for both the two-way MPA and the bootstrap CIs show that the nouns were mostly classified on the extremes (see Fig. 3, upper middle row). Something similar happened for the Imageability scale, where single nouns tended to populate the poles of the spectrum (see Fig. 3, lower middle row). These results agree with those found using CMDS. For the Concreteness scale, both criteria used to evaluate fit of the CMDS model revealed that the first two dimensions provided an adequate representation of the data (trace = 0.800; magnitude = 0.996). In the first dimension, the Concreteness scale seemed to create a dichotomy by separating the Abstract words from the remaining categories. No clear pattern could be identified in the second dimension. For the Imageability scale, both criteria used to evaluate fit of the CMDS model also revealed that the first two dimensions provided an adequate representation of the data (trace = 0.860; magnitude = 0.998). In the first dimension, the Imageability scale seemed to create a dichotomy by separating the Abstract words from the remaining categories. No clear pattern could be identified in the second dimension.

Results for the Multisensory scale suggested the presence of a stratified classification. This stratification could be observed in the results of both two-way MPA and bootstrap CIs. Instead, CMDS results suggested the presence of a dichotomization like the one observed for both the Concreteness and Imageability scales (see Fig. 3, lower row). The overall score for the two-way MPA was 6 and the analog R^2 for this model was 0.53, which represents a strong effect size. In the case of CMDS, both criteria revealed that the first two dimensions provided an adequate representation of the data (trace = 0.730; magnitude = 0.992). In the first dimension, the Multisensory scale seemed to create a dichotomy by separating the Abstract words from the remaining categories. The second dimension seemed to differentiate words in the extremes of the first dimension from words in the middle of the distribution.

In conclusion, the Action scale was the only instrument capable of identifying a continuous distribution of noun meaning. In contrast, the remaining scales tended to dichotomize the nouns or form stratified results.

Verbs

All rating scales created a continuous classification of the verbs. This could be observed in the results obtained by all three procedures (two-way MPA, 95% bootstrap CIs, and CMDS).

For the Action scale, the overall MPA score was 4.7 and the analog R^2 for the two-way fit was 0.43, which represents a moderate-to-strong effect size. The Action scale generated a continuous distribution of the verbs. These results agree with those obtained with 95% bootstrap CIs, since the verbs tended to span the entire spectrum of rating values (see Fig. 4, upper row). These results also agree with those found using CMDS. Both criteria revealed that the first two dimensions provided an adequate representation of the data (trace = 0.680; magnitude = 0.988). In the first dimension, words in the Action and Cognitive categories populated the extremes while the remaining terms were distributed between both poles, suggesting the presence of a continuum. The second dimension seemed more difficult to interpret since no pattern could be easily identified.

For the Concreteness scale, the overall MPA score was 5 and the analog R^2 for the two-way fit was 0.47, which represents a moderate-to-strong effect size. The Concreteness scale generated a continuous distribution of the verbs. These results agree with those obtained with 95% bootstrap CIs, since the verbs tended to span the entire spectrum of rating values (see Fig. 4, upper middle row). These results also agree with those found using CMDS. Both criteria revealed that the first two dimensions provided an adequate representation of the data (trace = 0.7; magnitude = 0.992). In the first dimension, words in the Action and Cognitive categories populated the extremes while the remaining terms were distributed between poles, suggesting the creation of a continuum. The second dimension seemed more difficult to interpret since no pattern could be easily identified.

For the Imageability scale, the overall MPA score was 5 and the analog R^2 for the two-way fit was 0.53, which represents a strong effect size. The Imageability scale generated a continuous distribution of the verbs. These results agree with those obtained with 95% bootstrap CIs, since the verbs tended to span the entire spectrum of values (see Fig. 4, lower middle row). These results also agree with those found using CMDS. Both criteria revealed that the first two dimensions provided an adequate representation of the data (trace = 0.720; magnitude = 0.993). In the first dimension, words in the Action and Cognitive categories populated the extremes while the remaining terms were distributed between poles, suggesting the presence of a continuum. The second dimension seemed more difficult to interpret since no pattern could be easily identified.

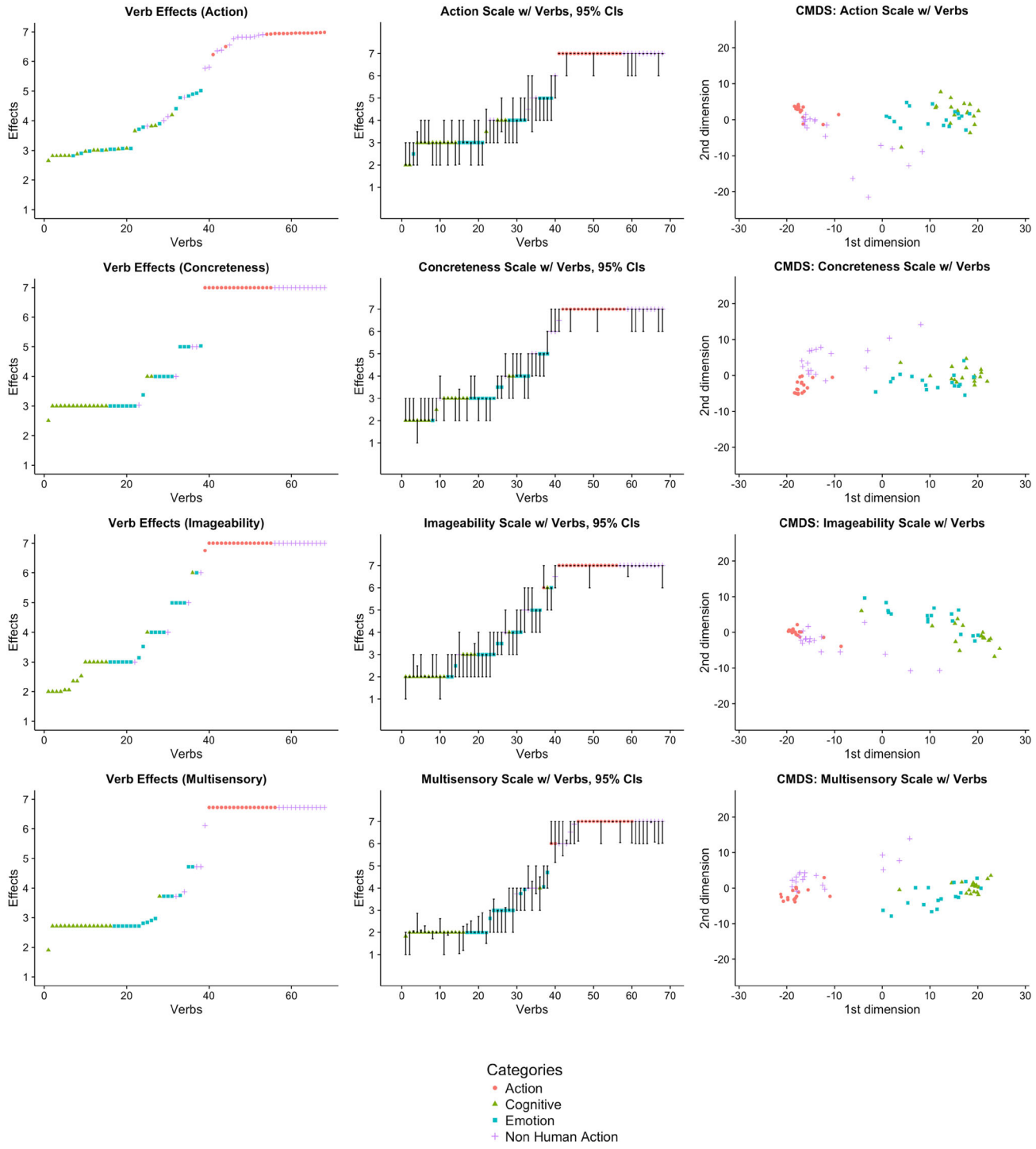

Fig. 4 Results from the two-way MPA models (*left column*), 95% bootstrap CIs (*middle column*), and CMDS models (*right column*) with Verbs. The *upper row* shows the results for the Action Scale; the *middle upper row* shows the results for the Concreteness scale; the *lower middle row* shows the results for the Imageability scale; the *lower row* shows the results for the Multisensory scale. Each word has been colored and shaped according to its preassigned Category

For the Multisensory scale, the overall MPA score was 4.5 and the analog R^2 for the two-way fit was 0.53, which represents a strong effect size. The Multisensory scale generated a continuous distribution of the verbs. These results agree with those obtained with 95% bootstrap CIs, since the verbs tended to span the entire spectrum of values (see Fig. 4, lower row). These results also agree with those found using CMDS. Both criteria revealed that the first two

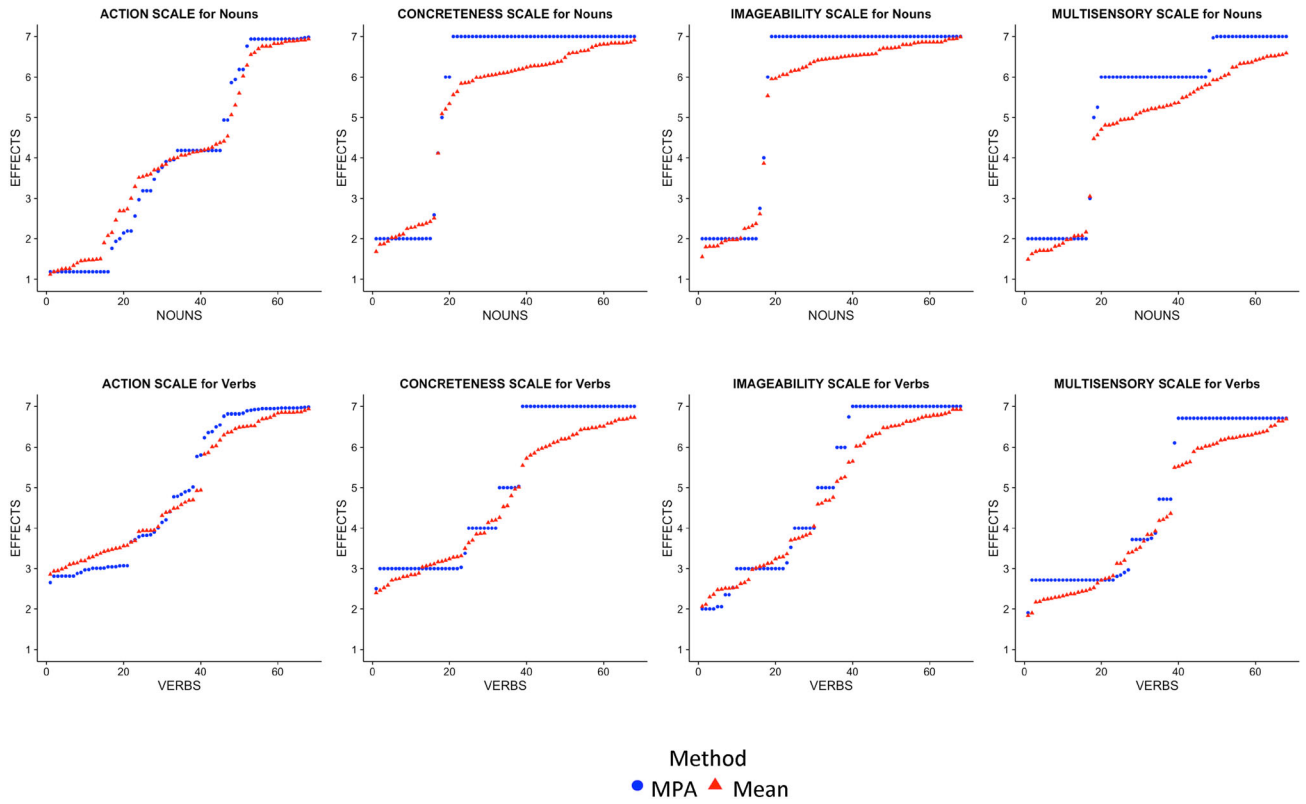

Fig. 5 Comparisons between MPA and mean results across the eight Experimental Blocks. MPA results are shown in *blue circles* and results for the mean analysis are shown in *red triangles*. In all figures, the mean tends to overemphasize the continuous distribution of word meaning. Using the mean may result in continua irrespective of the scale or stimuli used, implying that the statistical technique is driving the results. Conversely, MPA tends to identify continua under certain circumstances and to identify dichotomies in others. In this sense, MPA is sensitive to different scales and stimuli

dimensions provided an adequate representation of the data (trace = 0.720; magnitude = 0.993). In the first dimension, words in the Action and Cognitive categories populated the extremes while the remaining terms were distributed between poles, suggesting the creation of a continuum. The second dimension seemed more difficult to interpret since no pattern could be easily identified.

In summary, all rating scales identified a continuous distribution of verb meaning. Importantly, the Action scale was the only scale capable of doing so for both nouns and verbs.

Comparing MPA with sample mean analysis

To compare the performance of MPA against previously used methods based on means (Nelson & Schreiber, 1992; Wiemer-Hastings et al., 2001), the mean value for each word was calculated. Subsequently, we sorted and plotted each word′s mean value against the MPA results.

Figure 5 (upper row) shows each noun's mean value in the four rating scales together with the MPA estimates. Whenever the sample mean was used, a more continuous distribution was revealed. This situation was more pronounced for both the Action and Multisensory scales. In contrast, MPA only identified a clear continuum in the case of the Action scale. The same figure shows each verb's mean value in the four rating scales. Here, a continuum of word meaning was revealed for all rating scales when using both techniques. However, the sample mean seemed to overemphasize the continuous distribution of word meaning in the case of the Concreteness, Imageability, and Multisensory scales.

These results suggest that analyses based on means tend to result in continua irrespective of the scale and stimuli used. They do this by aggregating dissimilar responses into single scores containing multiple decimal values. In our view, this does not adequately represent the ordinal data obtained from the rating scales and can lead to an erroneous establishment of a continuum of word meaning.

Discussion

The current project suggests the use of a new procedure for analyzing word meaning data acquired through rating scales. We compared a robust method called MPA

(Hoaglin et al., 2000, 2006; Tukey, 1977) to standard approaches favoring the use of the sample mean. The work presented here advocates for the use of methodologies (MPA and bootstrap CIs) that have not been employed in this context in earlier studies, and that produce novel results that may change the way we investigate word meaning representation.

Using four different rating scales (Action, Concreteness, Imageability, Multisensory) and two word classes (Nouns, Verbs), we revealed that MPA can render important information along several dimensions. For instance, using three-way MPA models, we showed differences among rating scales in terms of how participants followed their instructions and how word categories were classified according to them (see Figs. 1, 2). In the same vein, using two-way MPA models, we showed that rating scales can identify continuous or dichotomous distributions of word meaning depending on the stimuli used (see Figs. 3, 4). Notably, our results demonstrated that MPA models are sensitive to different scenarios (stimulus type and rating scale) and do not always find continuous distributions of meaning as does the sample mean analysis. This means that MPA is not bound to identify word meaning continua, unless there is strong evidence of their presence in the data. This is in remarkable contrast to widely used methodologies based on the sample mean, whose implementation results in continua irrespective of the scale and stimuli used (see Fig. 5). This is because the mean tends to aggregate dissimilar responses, some of them outliers, into single scores that end up artificially discriminating among words.

Moreover, we presented further benefits of assessing rating scores with two-way MPA models, as they can account for both the words and participants involved in studies. In this context, MPA better classifies words than analyses based on means by assessing differences in criteria among participants and focusing on main tendencies within words, something that has been ignored by previous studies. This increase in performance is due to MPA's robustness (Hoaglin et al., 2000, 2006; Tukey, 1977), which neutralizes the effect of a good proportion of outliers. Remarkably, MPA does this without introducing the bias that would be present if we were to remove the outliers by hand.

We have also presented how the MPA procedure can easily provide 95% bootstrap CIs for each word. Using CIs diminishes the temptation to ascribe nonexistent differences among words and, therefore, reduces potential misinterpretations, at the same time that it represents valuable evidence for subsequent replication (Cumming, 2008, 2014). Additionally, we showed that we can evaluate the goodness of fit of an MPA model by calculating an analog R^2. This is an important step, as in doing so we get an estimation of how well the MPA model represents our data. The importance of calculating and reporting effect sizes has been greatly emphasized in later years, as doing so could improve the quality of psychological research (Cumming, 2014).

The idea of a concrete–abstract continuum is not new (Della Rosa, Catricalà, Vigliocco, & Cappa, 2010; Nelson & Schreiber, 1992; Wiemer-Hastings et al., 2001). Previous attempts to describe continua of word meaning used mean values to describe each word's position on rating scales (Nelson & Schreiber, 1992; Wiemer-Hastings et al., 2001). However, as shown here, evaluating the occurrence of continua via mean analysis masks the inherent variability present among participants. This is because the sample mean is not a robust measurement and is easily distorted by outliers. Although the mean is an appropriate descriptive for normally distributed symmetric data, it can be very non-robust for data like the rating scale data analyzed here. Furthermore, as we found in the current study, mean analyses create continua of word meaning irrespective of the instrument used. In the case of rating scores, this raises the question of whether the previously described continua were more a property of the statistical technique employed than a consequence of the adequacy of previously used rating scales.

Although this idea will need to be further addressed by future studies, our results provide preliminary evidence of how some rating scales dealt with different types of stimuli. Here, we compared the performance of four rating scales because of their popularity and/or because they represented distinct approaches on concrete and abstract word meaning representation. We selected the Concreteness and Imageability scales because of their widespread use (Altarriba et al., 1999; Paivio et al., 1968) and because some scholars have proposed that the difference between concrete and abstract terms occurs in terms of how easy or difficult it is to create a mental image of a word's referent (Paivio, 1986, 1991). Conversely, we selected the Multisensory scale after recent research highlighted the inappropriateness of reducing perceptual experience to visual imageability, as it happens with the Imageability scale (Connell & Lynott, 2012); and we selected the Action scale because it emphasizes the role of sensorimotor information in word meaning processing, an idea congruent with proponents of embodied accounts of meaning (Barsalou, 1999, 2008; Borghi et al., 2017).

Our results suggested that both the Concreteness and Imageability scales were prone to dichotomize noun meaning, find continuous distributions of verb meaning, and to assign higher values to nouns than to verbs. In contrast, the Multisensory scale was not prone to dichotomize nouns, but to create a stratified organization of them. The Multisensory scale also found continuous distributions of verb meaning and assigned higher values to nouns than to verbs. The

Action scale was prone to identify continuous distributions of nouns and verbs, but did not tend to assign higher values to nouns than to verbs, and vice versa. Taken together, these results suggest that the Action scale is a good instrument to differentiate among words, whether they are nouns or verbs. This also suggests that a rating scale based on how we act with word referents could be a key element in the classification of word meaning and that sensorimotor information could be a valuable component of the representations of word meaning. However, our results need to be considered with caution. This is because the words and categories included in our study were not exhaustive and, therefore, only represent a small sample of the potential words and categories that may be of interest to researchers.

The main goal of the current project was to present MPA as a tool to investigate word meaning data, but the steps followed here are not the only potential application for this robust procedure. For instance, consider a situation where a researcher wants to evaluate a set of words with similar properties across a large number of rating scales. This researcher wants to identify whether certain scales provide similar values to these words, because he/she is interested in whether these scales share some underlying properties. In this hypothetical scenario, this researcher could use a two-way MPA model with words organized along the *x*-axis and rating scales organized along the *y*-axis, and then explore the respective Word and Scale effects. Even more, if this researcher were to organize the words into categories, he/she could use a three-way MPA to estimate differences among categories by studying the corresponding Category effects. This example is only one of the potential applications of MPA, as the same could be said about any other data organized along two or more dimensions of interest.

Given the problems associated with the use of the sample mean for analyzing word meaning data, it is fundamental to seek new statistical practices. Here, we promote the use of MPA as a valuable and robust alternative, capable of successfully dealing with gross outliers. Additionally, we have shown that MPA can be integrated with the creation of bootstrap CIs and with the calculation of an effect size measure called analog R^2, two steps that are considered crucial to the improvement of psychological research (Cumming, 2008, 2014). Different from the sample mean analysis, MPA is not inclined to find continua in word meaning unless there is strong evidence of their existence. Because MPA can be a very useful statistical tool for researchers, we have included an R tutorial describing the same steps we have implemented in the current project (see Additional Materials). As pointed out before, this does not mean that our pipeline is the only potential application of MPA to word meaning data. Depending on the experimental design, MPA could be used for numerous purposes.

Acknowledgements The authors would like to thank Aesha Maniar and Devon Olson for their help with behavioral data collection. Felipe Munoz-Rubke would like to thank Fulbright and CONICYT for financial support provided through an academic scholarship.

Compliance with ethical standards

Conflict of interest Felipe Munoz-Rubke declares that he has no conflict of interest. Karen Kafadar declares that she has no conflict of interest. Karin H. James declares that she has no conflict of interest.

Research involving human participants All procedures performed in studies involving human participants were in accordance with the ethical standards of the Indiana University Bloomington Institutional Review Board, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all participants included in the study.

References

Allen, R., & Hulme, C. (2006). Speech and language processing mechanisms in verbal serial recall. *Journal of Memory and Language, 55*(1), 64–88. doi:10.1016/j.jml.2006.02.002.

Altarriba, J., Bauer, L. M., & Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society Inc, 31*(4), 578–602.

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. doi:10.3758/s13428-012-0215-z.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*(3), 445–459.

Barsalou, L. W. (1999). Perceptual symbol systems. *The Behavioral and Brain Sciences, 22*(4), 577–609. **(discussion 610–660)**.

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

Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience, 17*(6), 905–917.

Borghi, A. M., & Binkofski, F. (2014). *Words as social tools: An embodied view on abstract concepts*. New York: Springer.

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin, 143*(3), 263–292. doi:10.1037/bul0000089.

Borghi, A. M., Flumini, A., Cimatti, F., Marocco, D., & Scorolli, C. (2011). Manipulating objects and telling words: A study on concrete and abstract words acquisition. *Frontiers in Psychology*. doi:10.3389/fpsyg.2011.00015.

Borghi, A. M., Scorolli, C., Caligiore, D., Baldassarre, G., & Tummolini, L. (2013). The embodied mind extended: Using words as social tools. *Frontiers in Psychology*. doi:10.3389/fpsyg.2013.00214.

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

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM, 58*(3), 1–37. doi:10.1145/1970392.1970395.

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

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

Crutch, S. J., Troche, J., Reilly, J., & Ridgway, G. R. (2013). Abstract conceptual feature ratings: The role of emotion, magnitude, and other cognitive domains in the organization of abstract conceptual knowledge. *Frontiers in Human Neuroscience*. doi:10.3389/fnhum.2013.00186.

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 3*(4), 286–300. doi:10.1111/j.1745-6924.2008.00079.x.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. doi:10.1177/0956797613504966.

de Groot, A. M. (1989). Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(5), 824–845. doi:10.1037/0278-7393.15.5.824.

Della Rosa, P. A., Catricalà, E., Vigliocco, G., & Cappa, S. F. (2010). Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods, 42*(4), 1042–1048. doi:10.3758/BRM.42.4.1042.

Everitt, B., & Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. New York: Springer.

Fiebach, C. J., & Friederici, A. D. (2004). Processing concrete words: fMRI evidence against a specific right-hemisphere involvement. *Neuropsychologia, 42*(1), 62–70.

Fliessbach, K., Weis, S., Klaver, P., Elger, C. E., & Weber, B. (2006). The effect of word concreteness on recognition memory. *NeuroImage, 32*(3), 1413–1421. doi:10.1016/j.neuroimage.2006.06.007.

Ghio, M., Vaghi, M. M. S., & Tettamanti, M. (2013). Fine-grained semantic categorization across the abstract and concrete domains. *PLoS One, 8*(6), e67090. doi:10.1371/journal.pone.0067090.

Giesbrecht, B., Camblin, C. C., & Swaab, T. Y. (2004). Separable effects of semantic priming and imageability on word processing in human cortex. *Cerebral Cortex (New York, N.Y.: 1991), 14*(5), 521–529. doi:10.1093/cercor/bhh014.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(Supplement 1), 5228–5235. doi:10.1073/pnas.0307752101.

Hoaglin, D., Mosteller, F., & Tukey, J. W. (Eds.). (2000). *Understanding robust and exploratory data analysis* (Wiley classics library ed.). New York: Wiley.

Hoaglin, D., Mosteller, F., & Tukey, J. W. (Eds.). (2006). *Exploring data tables, trends, and shapes*. Hoboken: Wiley-Interscience.

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. doi:10.1016/j.neuropsychologia.2012.11.009.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '99) (pp. 50–57). New York, NY, USA:ACM Press. doi:10.1145/312624.312649.

Hubert, M., & Debruyne, M. (2009). Breakdown value. *Wiley Interdisciplinary Reviews: Computational Statistics, 1*(3), 296–302. doi:10.1002/wics.34.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*(12), 1217–1218. doi:10.1111/j.1365-2929.2004.02012.x.

Jessen, F., Heun, R., Erb, M., Granath, D. O., Klose, U., Papassotiropoulos, A., & Grodd, W. (2000). The concreteness effect: Evidence for dual coding and context availability. *Brain and Language, 74*(1), 103–112. doi:10.1006/brln.2000.2340.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1–37. doi:10.1037/0033-295X.114.1.1.

Kiefer, M., & Barsalou, L. W. (2013). Grounding the human conceptual system in perception, action, and internal states. In W. Prinz, M. Beisert, & A. Herwig (Eds.), *Action Science: Foundations of an Emerging Discipline* (pp. 381–407). Cambridge, MA:MIT Press.

Kloke, J., & McKean, J. W. (2015). *Nonparametric statistical methods using R*. Boca Raton: CRC Press, Taylor & Francis.

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. doi:10.1037/a0021446.

Kroll, J. F., & Merves, J. S. (1986). Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(1), 92–107. doi:10.1037/0278-7393.12.1.92.

Landauer, T. K. (1999). Latent semantic analysis: A theory of the psychology of language and mind. *Discourse Processes, 27*(3), 303–310. doi:10.1080/01638539909545065.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240. doi:10.1037/0033-295X.104.2.211.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *The Behavioral and Brain Sciences, 22*(1), 1–38. **(discussion 38–75)**.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2), 203–208. doi:10.3758/BF03204766.

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*(1–3), 59–70. doi:10.1016/j.jphysparis.2008.03.004.

Nelson, D. L., & Schreiber, T. A. (1992). Word concreteness and word structure as independent determinants of recall. *Journal of Memory and Language, 31*(2), 237–260. doi:10.1016/0749-596X(92)90013-N.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxfort: Oxford University Press.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 45*(3), 255–287. doi:10.1037/h0084295.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76*(1, Pt.2), 1–25. doi:10.1037/h0025327.

Papagno, C., Fogliata, A., Catricalà, E., & Miniussi, C. (2009). The lexical processing of abstract and concrete nouns. *Brain Research, 1263,* 78–86. doi:10.1016/j.brainres.2009.01.037.

Papagno, C., Martello, G., & Mattavelli, G. (2013). The neural correlates of abstract and concrete words: Evidence from brain-damaged patients. *Brain Sciences, 3*(3), 1229–1243. doi:10.3390/brainsci3031229.

Peirce, J.W. (2007). PsychoPy–Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1-2), 8–13. doi:10.1016/j.jneumeth.2006.11.017.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Recchia, G., & Jones, M. N. (2012). The semantic richness of abstract concepts. *Frontiers in Human Neuroscience*. doi:10.3389/fnhum.2012.00315.

Rodríguez-Ferreiro, J., Gennari, S. P., Davies, R., & Cuetos, F. (2011). Neural correlates of abstract verb processing. *Journal of Cognitive Neuroscience, 23*(1), 106–118. doi:10.1162/jocn.2010.21414.

Romani, C., Mcalpine, S., & Martin, R. C. (2008). Concreteness effects in different tasks: Implications for models of short-term memory. *The Quarterly Journal of Experimental Psychology, 61*(2), 292–323. doi:10.1080/17470210601147747.

Sabsevitz, D. S., Medler, D. A., Seidenberg, M., & Binder, J. R. (2005). Modulation of the semantic system by word imageability. *NeuroImage, 27*(1), 188–200. doi:10.1016/j.neuroimage.2005.04.012.

Schwanenflugel, P. J., Akin, C., & Luh, W.-M. (1992). Context availability and the recall of abstract and concrete words. *Memory & Cognition, 20*(1), 96–104. doi:10.3758/BF03208259.

Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(2), 344–354. doi:10.1037/0278-7393.14.2.344.

Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(1), 82–102. doi:10.1037/0278-7393.9.1.82.

Sidhu, D. M., Kwan, R., Pexman, P. M., & Siakaluk, P. D. (2014). Effects of relative embodiment in lexical and semantic processing of verbs. *Acta Psychologica, 149,* 32–39. doi:10.1016/j.actpsy.2014.02.009.

Spence, I., & Lewandowsky, S. (1989). Robust multidimensional scaling. *Psychometrika, 54*(3), 501–513. doi:10.1007/BF02294632.

Sullivan, G. M., & Artino, A. R. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education, 5*(4), 541–542. doi:10.4300/JGME-5-4-18.

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. doi:10.3758/BRM.40.4.1075.

Troche, J., Crutch, S., & Reilly, J. (2014). Clustering, hierarchical organization, and the topography of abstract and concrete nouns. *Frontiers in Psychology*. doi:10.3389/fpsyg.2014.00360.

Tsai, P.-S., Yu, B. H.-Y., Lee, C.-Y., Tzeng, O. J.-L., Hung, D. L., & Wu, D. H. (2009). An event-related potential study of the concreteness effect between Chinese nouns and verbs. *Brain Research, 1253,* 149–160. doi:10.1016/j.brainres.2008.10.080.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley Pub. Co.

Tyler, L. K., Russell, R., Fadili, J., & Moss, H. E. (2001). The neural representation of nouns and verbs: PET studies. *Brain: A Journal of Neurology, 124,* 1619–1634.

Varela, F. J., Thompson, E., & Rosch, E. (2000). *The embodied mind: Cognitive science and human experience* (8th ed.). Cambridge, Mass.: MIT Press.

Venables, W. N., & Ripley, B. D. (2010). *Modern applied statistics with S* (4th ed.). New York: Springer.

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. New York: Springer.

Wiemer-Hastings, K., Krug, J. D., & Xu, X. (2001). Imagery, context availability, contextual constraint and abstractness. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society* (pp. 1106–1111). Mahwah: Lawrence Erlbaum Associates Inc.