



How to carry out conceptual properties norming studies as parameter estimation studies: Lessons from ecology

Enrique Canessa^{1,2} · Sergio E. Chaigneau^{1,3} · Rodrigo Lagos⁴ · Felipe A. Medina^{4,5}

© The Psychonomic Society, Inc. 2020

Abstract

Conceptual properties norming studies (CPNs) ask participants to produce properties that describe concepts. From that data, different metrics may be computed (e.g., semantic richness, similarity measures), which are then used in studying concepts and as a source of carefully controlled stimuli for experimentation. Notwithstanding those metrics' demonstrated usefulness, researchers have customarily overlooked that they are only point estimates of the true unknown population values, and therefore, only rough approximations. Thus, though research based on CPN data may produce reliable results, those results are likely to be general and coarse-grained. In contrast, we suggest viewing CPNs as parameter estimation procedures, where researchers obtain only estimates of the unknown population parameters. Thus, more specific and fine-grained analyses must consider those parameters' variability. To this end, we introduce a probabilistic model from the field of ecology. Its related statistical expressions can be applied to compute estimates of CPNs' parameters and their corresponding variances. Furthermore, those expressions can be used to guide the sampling process. The traditional practice in CPN studies is to use the same number of participants across concepts, intuitively believing that practice will render the computed metrics comparable across concepts and CPNs. In contrast, the current work shows why an equal number of participants per concept is generally not desirable. Using CPN data, we show how to use the equations and discuss how they may allow more reasonable analyses and comparisons of parameter values among different concepts in a CPN, and across different CPNs.

Keywords Conceptual properties norming studies · Property listing task · Parameter estimation · Sample size determination · Sample coverage

Introduction

Researchers interested in studying concepts often characterize them by their properties (e.g., Schyns, Goldstone, & Thibaut,

1998) and their respective frequency distributions (Ashby & Alfonso-Reese, 1995; Griffiths, Sanborn, Canini, & Navarro, 2008; Rosch & Mervis, 1975). Properties may have continuous values (e.g., "height"; Goldstone, 1994; Tversky & Hutchinson, 1986), but a far more common practice is to treat them as binary (i.e., they either belong to a concept or not, e.g., the property "has four legs" may be a property of *dog* but not of *spider*). To study concepts, particularly those coded in language (e.g., *dog*), researchers often use the Property Listing Task (PLT). In the PLT, people are asked to produce semantic content for a given concept (e.g., for *dog*, people may produce "has four legs"). This content needs to be coded into properties that group verbalizations which differ only superficially into a single code (e.g., "has four legs" and "is a quadruped" might be coded as "four legs"). In what follows, we will refer to these coded properties simply as properties.

The PLT is widely used across psychology, both in basic and applied research (e.g., Hough & Ferraris, 2010; Perri, Zannino, Caltagirone, & Carlesimo, 2012; Walker & Hennig, 2004; Wu & Barsalou, 2009). Rather than studying

✉ Enrique Canessa
ecanessa@uai.cl

- ¹ Center for Cognition Research (CINCO), School of Psychology, Universidad Adolfo Ibáñez, Av. Presidente Errázuriz 3328, Las Condes, Santiago, Chile
- ² Faculty of Engineering and Science, Universidad Adolfo Ibáñez, Av. P. Hurtado 750, Lote H, Viña del Mar, Chile
- ³ Center for Social and Cognitive Neuroscience, School of Psychology, Universidad Adolfo Ibáñez, Av. Presidente Errázuriz 3328, Las Condes, Santiago, Chile
- ⁴ Programa de Bioestadística, Escuela de Salud Pública, Universidad de Chile, Santiago, Chile
- ⁵ Instituto de Estadística, Facultad de Ciencias, Universidad de Valparaíso, Valparaíso, Chile

single concepts, researchers are often interested in groups of concepts because they tend to organize themselves into semantic clusters. In conceptual properties norming studies (CPNs), PLT data are collected for a large set of concepts across many participants (e.g., Devereux, Tyler, Geertzen, & Randall, 2014; Kremer & Baroni, 2011; Lenci, Baroni, Cazzolli, & Marotta, 2013; McRae, Cree, Seidenberg, & Mcnorgan, 2005; Montefinese, Ambrosini, Fairfield, & Mammarella, 2013; Vivas, Vivas, Comesaña, García Coni, & Vorano, 2017). These norms can be represented as matrices containing different concepts with their respective properties' frequency distributions. Researchers use CPN data in at least two different ways. First, CPNs provide information about the underlying semantic structure of a representative individual (e.g., showing that, on average, *dog* and *cat* are conceptually more similar to each other than either is to *cup*), thus allowing researchers to test theories about the nature of concepts and conceptual content (e.g., Cree & McRae, 2003; Rosch & Mervis, 1975; Vigliocco, Vinson, Lewis, & Garrett, 2004; Wu & Barsalou, 2009). Second, CPNs may be used as a source of normed stimuli and of control variables for experiments (McRae, Cree, Westmacott, & De Sa, 1999; Bruffaerts, De Deyne, Meersmans, Liuzzi, Storms, & Vandenberghe, 2019).

As is customary in the field, we acknowledge that conceptual properties obtained in CPNs are not equivalent to the underlying organization of semantic memory. Rather, they are generally thought to provide a window into semantic memory, to which we do not have direct access (e.g., McRae, Cree, Seidenberg & Mcnorgan, 2005). CPN data are verbal properties, while the underlying properties are in some unknown format and some of them may even be not verbalizable (e.g., faces may be difficult to describe by using words, but they are nonetheless characterized by features that can be used for recognition and categorization). However, a fact that is often overlooked is that verbalizable semantic properties are important in their own right because they are the kind of data that we do have access to, and that these data may be incomplete, not only because they may not accurately reflect the underlying semantic structures, but because the population of potential properties may not be appropriately sampled, which will be the focus of the current work.

Their usefulness notwithstanding, a previously unacknowledged feature of CPNs as a research strategy is that values obtained from CPN data are routinely treated as population parameters rather than as parameter estimates. Here, it is useful to recall that, given that on any study we generally do not have access to the entire population of interest, but only to a representative sample, we are forced to estimate the true unknown parameters of the entire population. Thus, when researchers use a CPN to obtain normed stimuli, or when they use values associated with concepts and properties for control purposes, those raw values are at best unbiased point

estimations of the true population values, not the population values themselves. As we will discuss in short, this poses a problem that may surface in at least three different but related guises: the issue of generalizing results based on data from norms (CPNs and other similar norms), the issue of deciding on sample sizes for CPNs, and the issue of replicability of CPNs.

Reasons for overlooking that metrics obtained from CPN data are in fact parameter estimates, may stem in part from not taking into account the difference between a study's internal and external validities. Recall that internal validity is achieved when an experiment is correctly designed, so that its conclusions logically follow from its methods (it is related to experimental design). External validity, in contrast, has to do with whether data are representative of the population (it relates to sample representativeness). We speculate that researchers that collect CPNs come from an experimentalist tradition where internal consistency, and not representativeness, is the basic criterion to decide about generalizability (though bear in mind that CPNs are not experiments). However, we believe that the most influential factor for not treating CPNs as parameter estimation studies derives from a tradition that focuses almost exclusively on CPNs as a means of measurement (i.e., measuring similarity and other such cognitive constructs). To the best of our knowledge, this tradition can be traced back to Eleanor Rosch's now classical studies (Rosch & Mervis, 1975). In fact, Rosch and Mervis computed similarity as the number of shared properties weighted by their frequencies, but weeded out of their data properties with frequency 1 (i.e., those reported by a single participant), precisely on grounds that singletons did not contribute to the measurement of the overall similarity structure (i.e., properties with frequency 1 were viewed as measurement error). From that point on, the practice of weeding-out low-frequency properties from CPNs (typically, frequencies lower than 5) on grounds of reducing measurement error has been continued in most, if not all, CPN studies, thus supporting our conclusion that CPNs have been conceptualized as a measurement procedure (e.g., Ashcraft, 1978; Coley, Hayes, Lawson, & Moloney, 2004; Cree & McRae, 2003; Devereux, Tyler, Geertzen, & Randall, 2014; Garrard, Lambon Ralph, Hodges, & Patterson, 2001; Hampton, 1979; Kremer & Baroni, 2011; Lenci, Baroni, Cazzolli, & Marotta, 2013; McRae, Cree, Seidenberg, & Mcnorgan, 2005; Vinson & Vigliocco, 2008).

Note that if participants producing properties for a given concept shared their conceptual content to a high degree, then interpreting low-frequency properties as noise could be warranted. However, there are two related reasons that suggest this is not a good practice. First, there is evidence that there is substantial variability in the PLT data between individuals, and within the same individual across time (Barsalou, 1987; Chaigneau, Canessa, Barra, & Lagos, 2018). When feature overlap is researchers' main goal, and property frequency

distributions are pruned, part of this variability is lost (a similar argument in favor of retaining the long tails of property frequency distributions can be found in De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019). Second, even if shared properties are the focus (e.g., if one wants to determine the strongest properties that characterize and differentiate a specific concept from another one), disregarding a portion of those properties that are not shared, will produce an overestimation of the proportion of shared properties. Consider, e.g., not including low-frequency properties (whatever the chosen cut-off point is) when attempting to measure semantic distances, which would misrepresent distances, making them smaller than what the true value possibly is.

Yet another reason for paying attention to low-frequency properties is theoretical, rather than statistical. Note that weeding out low-frequency properties is equivalent to having a statistical definition of what a valid conceptual property is. A seldom noted fact is that the field does not have a formal definition of what a property is. For example, regarding abstract concepts (e.g., *democracy*), properties listed by participants are generally other concepts (e.g., “voting,” “president”), and the field’s current practices imply considering those concepts as properties. What we propose in the current work is that we do not need to continue using a statistical definition of semantic properties (i.e., only those above a certain frequency threshold are true properties). In fact, it may be misleading to do so.

As already anticipated, disregarding the parameter estimation perspective on CPNs poses a threefold problem. A first angle on this problem is the question of how to generalize results based on data from CPNs. Because metrics calculated from raw CPN data are only point estimations and not the population parameters themselves, it is questionable to what extent conclusions from studies using those estimates generalize. Take for example a reaction time (RT) study by Pexman, Hargreaves, Siakaluk, Bodner and Pope (2008). In that study, the authors obtained different measures of a construct called semantic richness (SR) for a given word, all collected in different norming studies, and regressed them on RT data from lexical and semantic decision tasks (respectively, LDT and SDT). For each word, these measures included the number of words co-occurring in similar lexical contexts (i.e., number of semantic neighbors), the distribution of a word’s occurrences across content areas (i.e., contextual dispersion), and the total number of unique properties listed for that word in a CPN (SR, hereinafter denoted by S_{obs}). From their results, the authors draw the general conclusion that because the three measures predict unique variance in LDT and SDT, and because the measures themselves are only modestly inter-correlated, they appear to tap on different constructs. However, because the best scenario is that the estimations obtained from the norming studies are equally likely to be under or overestimations of the true population values,

Pexman et al.’s study’s generalizability depends critically on the (unknown) quality of the SR estimators. Note that many other studies may be subject to the same considerations, given that other variables whose computation depends on having a good description of the property frequency distribution might suffer the same problem (e.g., property generality, cue validity, property specificity, semantic neighborhood density).

Aside from generating uncertainty about how to interpret results based on variables obtained from CPNs, as discussed above, there are two other related and important problems. Prior to data collection, researchers have to decide about sample size (i.e., how many participants will list properties for each concept). Though sample sizes vary, perusing the literature suggests that researchers have implicitly agreed that somewhere between 20 and 30 participants listing properties for a given concept is a reasonable number (e.g., Cree & McRae, 2003; Devereux, Tyler, Geertzen, & Randall, 2014; Lenci, Baroni, Cazzolli, & Marotta, 2013; McRae, Cree, Seidenberg, & Mcnorgan, 2005; Montefinese, Ambrosini, Fairfield, & Mammarella, 2013). However, note that, other than tradition, no explicit rationale is given for this decision. Furthermore, and perhaps due to these same considerations about measurement precision, some researchers assume that larger studies are ipso facto more precise (e.g., De Deyne et al., 2019), disregarding that larger studies are more likely to introduce error due to many different problems that affect census-type studies, e.g., changes in concept meaning in the population due to the possibly long period necessary for a census-type study (see, e.g., De Deyne et al., 2019 study, which took several years to complete); and errors in processing the huge amount of raw data collected, which needs to be curated and coded. Contrary to deciding this issue by consensus, by practical considerations, or merely by intuition, in our view sample size should be explicitly justified, as is routine in parameter estimation studies.

The third and final problem is the following. Surprisingly, there is currently no clear and agreed upon way to compare CPNs. Imagine two different CPNs collecting data for the same set of concepts from two different samples. When could we say that they are replications of each other? When can we meaningfully make comparisons across studies? One current way to make the comparison would be to test if there are statistically significant differences between studies’ results. To the extent that one finds non-significant differences, one could say that the CPNs are comparable (this is in part the approach taken in Kremer & Baroni, 2011). However, this procedure problematically assumes that researchers are aiming at finding null results. Another procedure that could be used to judge whether CPNs are replications would be to compute a similarity measure between concepts in two separate CPNs and then to test if those similarities are correlated across CPNs (e.g., *virus* and *bacteria* should be about as similar to each other in CPN 1 as they are in CPN 2). However,

because even small correlations can be significant given a sufficiently large sample (in this case, number of concepts in the datasets), a significant correlation is insufficient, and a somewhat arbitrary cut-off point has to be established to answer the question of whether the norms are similar enough to be considered replications. This is reminiscent of defining arbitrary cut-off values for reliability computations in psychometrics.

A further complication when comparing norms is that many decisions in norm collecting and data processing could affect obtained data, creating spurious differences that do not stem from population differences (we will return to this point later in the current work). To illustrate differences in data processing, consider procedures discussed in Buchanan, De Deyne, and Montefinese (in press), where instead of coding a whole phrase to obtain a conceptual property (as we describe in the current work), sentences are parsed such that each noun becomes a separate property in the norms, i.e., the bag-of-words approach. The reader may note that data processing also introduces questions about CPNs replicability; a problem that has been almost completely overlooked in the CPN literature, with few notable exceptions (Bolognesi, Pilgram, & van den Heerik, 2017).

As discussed above, sample size estimation and comparability problems stem from researchers treating data from CPNs as the population rather than as mere samples, something that seems related to the tradition of viewing these studies as procedures for measuring similarity and related constructs, but overlooking that researchers are in fact estimating population parameters. This analysis provides the goal for the current work. To be able to better handle issues of data interpretation, sample size and inter-studies comparability, cognitive researchers interested in CPN data would benefit from statistical methods that would allow them to treat CPNs as parameter estimation studies. Our current aim is to provide researchers that collect or use CPN data with appropriate statistical methods and guidelines to interpret them. To this end, we freely draw from work on the problem of species richness estimation in ecology (Chao & Chiu, 2016), which offers an extremely close parallel to problems found in CPN studies. In the rest of this paper, we discuss some estimators that can be imported into CPNs from the field of ecology, illustrate their usage by means of a locally obtained CPN, and discuss how researchers who collect and use CPN data may benefit from these methods.

Estimators

The estimators discussed here are more fully reviewed in Chao and Chiu (2016) in the context of problems of species richness estimation in ecology. However, as these same authors discuss, the problem is more general than that, and the

estimators are widely applicable in many different disciplines. In our particular case, the formulae we use in the current work allow using semantic richness (SR; Pexman, Hargreaves, Edwards, Henry, & Goodyear, 2007; Pexman, Hargreaves, Siakaluk et al., 2008; Recchia & Jones, 2012) defined as the total count of unique properties associated with a given concept in a given CPN (S_{obs}), to calculate an estimate (\hat{S}) of the corresponding population parameter (S). Because S_{obs} has been shown to predict processing speed in timed tasks that require cognitive effort (e.g., LDT, SDT), S_{obs} values computed from CPNs are of interest in cognitive research (Hargreaves & Pexman, 2014; Kounios, Green, Payne, Fleck, Grondin, & McRae, 2009). Even more importantly, the total number of properties obtained for a given concept will affect any other measure derived from CPN matrices because it is associated with the shape of the corresponding property frequency distribution. From here on, we will closely follow the notation used in Chao and Chiu (2016), so that the interested reader may easily examine that work.

To introduce the different values necessary to compute the estimator, as well as the assumptions being made, we resort to an idealized PLT study. Imagine a researcher who wants to estimate the number of conceptual properties (S) associated with a given concept and hence performs a single PLT (perhaps within a broader CPN study). To that end, her participants (T = number of participants producing properties for a given concept) list responses that are tokens of one and only one of the S_{obs} coded properties that correspond to the concept. After collecting her data, the researcher can arrange it in a property by participant incidence matrix with S_{obs} rows and T columns, where each matrix-cell (W_{ij}) contains a 1 if participant j produced property i (0 otherwise). Assuming that the i -th property has a constant incidence probability (π_i), i.e., that the probability that property i is produced by a participant is the same for all participants, each element w_{ij} in the incidence matrix is a realization of a Bernoulli random variable with success probability π_i (i.e., $P(W_{ij} = 1) = \pi_i$ and $P(W_{ij} = 0) = 1 - \pi_i$). Thus, the probability distribution for the incidence matrix can be expressed as:

$$P(W_{ij} = w_{ij}) = \prod_{i=1}^{S_{\text{obs}}} \prod_{j=1}^T \pi_i^{w_{ij}} (1-\pi_i)^{1-w_{ij}} = \prod_{i=1}^{S_{\text{obs}}} \pi_i^{y_i} (1-\pi_i)^{T-y_i} \quad (1)$$

, where y_i is the number of tokens of the i -th property that are observed in the sample (i.e., the property frequency in the marginal column of the $S_{\text{obs}} \times T$ matrix, $y_i = \sum_{j=1}^T W_{ij}$).

Generally, CPNs report y_i frequencies for each of the S_{obs} properties obtained for each of the concepts in the study (but remember that CPNs typically weed out low-frequency data). The marginal distribution for the incidence-based frequency $Y_i = y_i$ for the i -th property follows a binomial distribution:

$$P(Y_i = y_i) = \binom{T}{y_i} \pi_i^{y_i} (1-\pi_i)^{T-y_i} \tag{2}$$

Note that this model not only assumes that π_i is constant for each property, but also that properties are independent (i.e., that a property’s detectability is independent from other properties being or not detected). Furthermore, it assumes that the number of properties associated with a given concept in the population is finite. Some of these assumptions can be relaxed, but at the expense of making the application of the model more complicated. We will consider these issues in greater depth in the “The necessary simplifications” subsection. Note also that Eqs. (1) and (2) are a non-parametric model, which means that it does not need to assume a known probability distribution for the π_i ’s, making it even more general (Chao & Chiu, 2016).

Now, as we have argued above, our researcher is not interested in her particular sample of participants in itself, but rather on being able to estimate S and on having criteria to determine an appropriate sample size for her study. For the estimators we discuss in the current work, the only information needed from the incidence matrix are the number of properties reported by a single individual (Q_1) and the number of properties reported by only two individuals (Q_2) (respectively, the number of singletons and the number of doubletons). More generally, singletons and doubletons are two of the incidence-based frequency counts ($Q_0, Q_1, Q_2, \dots, Q_T$), where Q_k corresponds to the number of properties that are reported by exactly k participants, $k = 0, 1, \dots, T$. The unobserved Q_0 frequency count represents the number of properties not reported by any of the T participants.

The intuition behind being able to make the necessary estimation is simple. If a sample contains many singletons, then it is likely that there are still properties in the population that are not covered in the sample (assuming a finite number of properties in the population). However, once the sample starts to produce repetitions (only doubletons, tripletons, etc., i.e., participants have at least one common property among them), this signals that coverage is almost complete. Then, the value \widehat{S} estimates the true value for semantic richness by estimating Q_0 (i.e., $\widehat{S} = S_{\text{obs}} + \widehat{Q}_0$). Though we do not provide further mathematical details for the estimators below, the interested reader is referred to Chao and Chiu (2016) and references therein, where the full derivation of the expressions is presented. However, note that if the model stated in Eqs. (1) and (2) and its related simplifications are good approximations to reality (i.e., the model is valid), then all the following formulae are also valid, which were derived using the standard method of moments estimation and asymptotic approach (Chao & Chiu, 2016 and references therein).

Thus,

$$\widehat{S} = \begin{cases} S_{\text{obs}} + A \frac{Q_1^2}{2Q_2} & \text{if } Q_2 > 0 \\ S_{\text{obs}} + A \frac{Q_1(Q_1-1)}{2} & \text{if } Q_2 = 0 \end{cases} \tag{3}$$

, where $A = \frac{(T-1)}{T}$

And where the estimator’s variance can be approximated by Eq. (4).

$$\widehat{\text{var}}(\widehat{S}) = \begin{cases} Q_2 \left[\frac{A}{2} \left(\frac{Q_1}{Q_2} \right)^2 + A^2 \left(\frac{Q_1}{Q_2} \right)^3 + \frac{A^2}{4} \left(\frac{Q_1}{Q_2} \right)^4 \right] & \text{if } Q_2 > 0 \\ A \frac{Q_1(Q_1-1)}{2} + A^2 \frac{Q_1(2Q_1-1)^2}{4} - A^2 \frac{Q_1^4}{4\widehat{S}} & \text{if } Q_2 = 0 \end{cases} \tag{4}$$

From this variance, we can calculate the standard error of \widehat{S} (i.e., s.e. $\widehat{S} = \sqrt{\widehat{\text{var}}(\widehat{S})}$) and the confidence interval for S can be computed by Eq. (5)

$$95\%CI \text{ for } S = \left[S_{\text{obs}} + \frac{\widehat{S}-S_{\text{obs}}}{D}, S_{\text{obs}} + \left(\widehat{S}-S_{\text{obs}} \right) D \right] \tag{5}$$

, where

$$D = \exp \left[1.96 \sqrt{\ln \left(1 + \frac{\widehat{\text{var}}(\widehat{S})}{\left(\widehat{S}-S_{\text{obs}} \right)^2} \right)} \right] \tag{6}$$

Note that using Eq. (5) assumes that $\ln(\widehat{S}-S_{\text{obs}})$ is approximately normally distributed, i.e., $(\widehat{S}-S_{\text{obs}})$ follows an approximate log-normal distribution. Also see that $(\widehat{S}-S_{\text{obs}})$ corresponds to the rightmost summand in Eq. (3).

Estimating a desirable sample size requires estimating how well the current sample covers the true number of properties associated with the given concept in the population, where coverage is defined in general terms as the fraction of the total number of properties in the population that are captured in the total sample of T participants. More formally, coverage is defined as the fraction of the total incidence probabilities of the reported properties that are in the reference sample (Chao & Chiu, 2016). Coverage can be estimated by Eq. (7):

$$\widehat{C}(T) = 1 - \frac{Q_1}{U} \left[\frac{Q_1 (T-1)}{Q_1 (T-1) + 2Q_2} \right] \quad (7)$$

, where $U = \sum_{k=1}^T kQ_k = \sum_{i=1}^{S_{\text{obs}}} Y_i$ (i.e., total number of properties listed by all participants for a concept).

Furthermore, the same logic used in deriving Eq. (7) allows estimating the coverage expected from increasing sample size in t^* participants.

$$\widehat{C}(T + t^*) = 1 - \frac{Q_1}{U} \left[\frac{Q_1 (T-1)}{Q_1 (T-1) + 2Q_2} \right]^{(t^*+1)} \quad 0 \leq t^* \leq 2T \quad (8)$$

From Eq. (8), we can solve for t^* and estimate the number of additional participants needed to obtain a certain target coverage ($\widehat{C}_{\text{target}}$):

$$t^* = \text{ceiling} \left[\frac{\ln \left(\frac{U}{Q_1} \left[1 - \widehat{C}_{\text{target}} \right] \right)}{\ln \left(\frac{(T-1)Q_1}{(T-1)Q_1 + 2Q_2} \right)} - 1 \right] \quad 0 < t^* \leq 2T \quad (9)$$

, where the *ceiling* function returns the closest integer that is larger than or equal to the corresponding argument of that function.

The expected value for \widehat{S} resulting from the same t^* increment of participants can be estimated by Eq. (10).

$$\widehat{S}(T + t^*) = S_{\text{obs}} + \widehat{Q}_0 \left[1 - \left(1 - \frac{Q_1}{T\widehat{Q}_0 + Q_1} \right)^{t^*} \right] \quad 0 < t^* \leq 2T \quad (10)$$

, where $\widehat{Q}_0 = \widehat{S} - S_{\text{obs}}$ and, as already noted, corresponds to the rightmost summand in Eq. (3).

For those familiarized with qualitative research, note that coverage equations offer a formalization of the idea of theoretical saturation (Glasser & Strauss, 1967). In the next section, we illustrate the use of the presented formulae to estimate SR using PLT data obtained from a local CPN.

CPN data

To apply the presented formulae, we need the number of singletons (Q_1) and doubletons (Q_2) produced by participants for each concept. Unfortunately, and as already discussed, those values are routinely dismissed from CPN data, and thus not reported. We made efforts to find CPN data reporting singletons and doubletons by contacting authors of reported CPNs.

However, no author could furnish us with the necessary data to compute those values within reasonable effort (i.e., without us having to recode all the raw output for every participant and concept). Therefore, we used our own norms, in which we collected data for abstract concepts for a previous study and where we processed all listed properties regardless of their frequencies.

In our norms, for each concept, participants were required to write down words or short phrases that would allow someone else to correctly guess the concept, following the procedure described in Recchia & Jones, 2012. Each participant ($N = 100$, all Chilean Spanish speakers), was asked to produce properties for ten randomly chosen concepts taken from a total list of 27 abstract concepts. Note that using this procedure allowed us to obtain a different number of participants for each concept, something that we make use of in the analyses that will be presented next. For each of the 27 concepts, we obtained properties from 36.6 participants on average (min = 22, max = 52), resulting in a total of 5457 token responses.

These token responses were coded as follows. During a first coding phase, a trained coder classified the 5457 responses as valid or invalid. Valid responses provided conceptual content (e.g., producing “money” for the concept *happiness*). Invalid responses were cue repetitions (e.g., producing “deciding what to put on my nightstand” as a response for the concept *decision*), property repetitions, and metacognitive (e.g., “this is hard”) or off-task comments. During a second coding phase, the coder grouped the 4941 remaining valid responses into 729 response types. To estimate the reliability of the coding process, a second coder received the 729 codes, and proceeded to independently recode the 4941 valid responses. Following Hallgren’s (2012) recommendation, we computed Cohen’s kappa (Cohen, 1960) as a reliability estimate. The advantage of using kappa instead of simply using the percentage of agreement, as is often done (Bolognesi, Pilgram, & van den Heerik, 2017), is that kappa corrects for chance. Coding produced a kappa of .76, which is considered a substantial agreement (Landis & Koch, 1977), suggesting that our subsequent analyses were not unduly affected by unreliability concerns.

As with any carefully collected CPN data, different metrics can be computed from our data. To show that our CPN is not different in that sense, we computed a pairwise 27 x 27 symmetric distance matrix and submitted it to a clustering algorithm. The chosen distance measure required obtaining the number of shared properties between all pairs of concepts, and using it to compute a Jaccard distance for each pair (Jaccard, 1901), defined as 1 minus the ratio between intersecting properties over the union of properties. (Note that Jaccard similarity is a special case of Tversky’s 1977 set-theoretic contrast similarity measure, when similarity is symmetric. Also note that we could have used any other distance measure, but that is irrelevant for the present study.) Then,

those distances were submitted to a Weighted Pair Group Method with Arithmetic Mean Hierarchical Clustering algorithm (Sokal & Michener, 1958). Figure 1 shows that the clustering algorithm was able to retrieve a similarity structure contained in our data, which is in itself interesting above and beyond the focus of the current work. For example, inspecting Fig. 1 shows that *success* and *happiness* are found close by in the semantic space, as so are *compassion* and *gratitude*, which, together with *hope*, all group together at a higher level to form a cluster of positive emotions (other clusters are similarly sensible).

Though it is not the focus of our current work, Fig. 1 suggests that we could successfully use our CPN's data to guide experiments, which is precisely one of the uses of CPNs. Two examples should suffice to illustrate. If we were to conduct a lexical decision task with priming, we could reasonably predict that *truth* would be a better prime for *honesty* than *excuse*, given that the former is much more similar to *honesty* than the latter. If we were to look for content differences in properties being produced, we could group concepts by valence using our clustering solution (i.e., negative emotionality, positive emotionality, neutral emotionality) and look for differences between those groups. In that sense, our CPN is very comparable to most other published CPNs, both in its results and in its potential utility. Additionally, note that using a CPN for abstract concepts (as we do here), is irrelevant for the purpose of the present work, given that Eqs. (1) to (10) apply regardless of the type of concept used (i.e., concrete or abstract concepts).

Using the estimators

To obtain the estimators discussed in the corresponding section, we computed the number of singletons (Q_1) and

doubletons (Q_2) from the CPN data, as well as S_{obs} (the number of unique properties) and the total number of properties produced (U), for each of the 27 concepts included in our CPN. Then, using the number of participants who listed properties for each concept (T) and Eqs. (3) to (7), we calculated the estimated value for the total number of properties that describe a concept (\hat{S}), the standard error of that figure, the corresponding 95% confidence interval for S and the estimated coverage (\hat{C}) reached by the number of participants who listed properties for each concept. Table 1 presents all of these figures.

As can be observed in Table 1, estimated sample coverages ($\hat{C}(T)$) for the 27 concepts are only modest, ranging from 54% to 78%, suggesting that there is substantial information that is not being captured in our data. Differences in coverage are important because it is questionable whether comparing concepts with different coverages (e.g., comparing semantic richness values, i.e., S_{obs}) makes sense when the observed values are only rough estimations of the true values. Importantly, note that the different coverages are not only a function of different sample sizes (T), but most importantly of the properties' frequency distribution. In fact, a visual inspection of Table 1 shows that the values of Q_1 and Q_2 vary noticeably among concepts, indicating shorter and longer tails of the corresponding property frequency distributions, where concepts with shorter tails (i.e., larger Q_2 values and smaller Q_1 values) exhibit higher coverages than concepts with longer tails (i.e., smaller Q_2 values and larger Q_1 values). Thus, solely increasing T within reasonable bounds does not guarantee good coverage.

This same information can be easily appreciated in Fig. 2, which shows the point estimates of semantic richness (i.e., \hat{S}) and corresponding 95% CI. Note that given that $\hat{S}-S_{\text{obs}}$ follows a log-normal distribution with long asymmetric tails, the CIs are not symmetric around the point estimate \hat{S} .

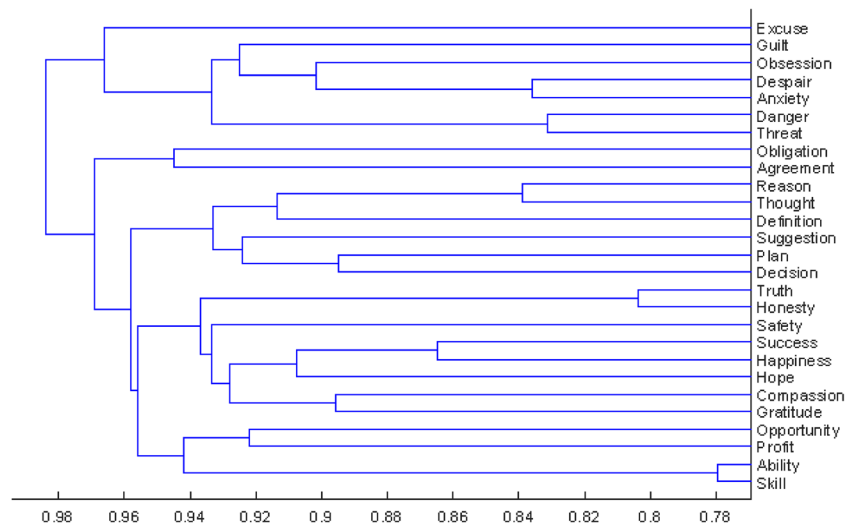


Fig. 1 Weighted linkage dendrogram for 27 abstract concepts

Table 1 Results from applying Eqs. (3) to (7) to CPN data for calculating \hat{S} , std. error of \hat{S} , 95% CI for S , and calculated \hat{C} ; and Eqs. (9) and (10) for calculating t^* and $\hat{S}(T + t^*)$

Concept	Q_1	Q_2	T	S_{obs}	\mathcal{U}	\hat{S}	s.e. \hat{S}	95% CI for S	$\hat{C}(T)$	t^*	$\hat{S}(T + t^*)$
Ability	45	7	27	64	124	203.3	68.1	120.2 409.1	64%	42	117.8
Agreement	56	15	46	89	198	191.3	39.3	138.4 300.7	72%	21	110.6
Anxiety	72	16	42	108	214	266.1	55.8	188.8 417.4	67%	39	161.2
Compassion	49	8	31	71	152	216.2	67.1	132.3 414.9	68%	35	120.0
Danger	74	17	51	106	256	263.9	54.5	187.8 410.8	71%	29	141.7
Decision	53	9	24	70	117	219.6	65.7	135.6 410.9	55%	49	145.6
Definition	45	7	30	64	139	203.8	68.3	120.4 410.4	68%	36	107.6
Despair	62	15	40	96	189	220.9	46.6	157.6 349.5	68%	32	135.7
Excuse	52	8	29	68	116	231.2	74.4	137.6 450.6	56%	65‡	150.1‡
Gratitude	45	11	51	76	201	166.2	39.4	115.8 280.7	78%	0	76.0
Guilt	67	8	35	87	164	359.5	118.3	207.7 702.3	59%	88‡	210.3‡
Happiness	65	11	37	92	210	278.9	74.2	180.2 487.7	69%	36	144.2
Honesty	49	12	37	74	179	171.3	40.7	118.3 288.0	73%	16	91.9
Hope	64	16	40	96	195	220.8	45.5	158.4 345.5	68%	31	135.6
Obligation	45	10	32	64	150	162.1	43.8	106.6 290.1	70%	21	88.3
Obsession	51	11	28	74	121	188.0	48.1	125.6 326.1	59%	41	127.5
Opportunity	56	8	31	73	129	262.7	85.2	154.9 512.4	57%	71‡	165.0‡
Plan	69	18	52	109	241	238.7	45.2	175.8 360.7	72%	25	137.1
Profit	34	14	33	62	145	102.0	18.5	78.9 156.9	77%	2	64.0
Reason	70	13	37	97	188	280.4	68.5	187.3 469.3	63%	51	170.6
Safety	99	18	48	136	275	402.6	84.2	281.7 623.8	64%	63	237.3
Skill	53	11	43	81	205	205.7	52.1	137.8 354.7	74%	16	98.1
Success	55	8	30	77	156	259.8	82.4	155.7 501.6	65%	47	144.3
Suggestion	41	8	22	57	109	157.3	48.4	97.9 302.7	63%	29	97.4
Thought	79	13	46	112	235	346.8	85.3	229.8 580.0	67%	58	190.7
Threat	57	8	39	80	175	277.9	88.5	165.6 537.1	68%	53	141.9
Truth	53	11	26	70	114	192.8	51.3	125.9 339.5	54%	45	133.3

‡ = t^* exceeds $2T$, thus actual value of t^* should be $2T$: Excuse $t^* = 58$ $C(T + t^*) = 76.5\%$ $S(T + t^*) = 144.6$, Guilt $t^* = 70$ $C(T + t^*) = 75.1\%$ $S(T + t^*) = 192.6$, Opportunity $t^* = 62$ $C(T + t^*) = 76.1\%$ $S(T + t^*) = 157.3$

More formally returning to the effect that T and property frequency distribution have on coverage, we can replace \mathcal{U} in Eq. (7) by $\sum_{k=1}^T kQ_k = Q_1 + 2Q_2 + 3Q_3 + \dots$ and by doing the multiplication with the term in square brackets, obtain:

$$\hat{C}(T) = 1 - \frac{Q_1^2 (T-1)}{(T-1)(Q_1^2 + 2Q_1Q_2 + 3Q_1Q_3 + \dots) + 2Q_1Q_2 + 4Q_2^2 + 6Q_2Q_3 + \dots} \quad (11)$$

Expression (11) more explicitly shows the interplay between sample size (T), the frequency distribution of sampled properties (the Q_k variables) and the interaction among the Q_k variables. From Eq. (11), note that an increase in Q_2, Q_3, \dots, Q_T (i.e., an increase in all Q_k with $k \geq 2$), increases coverage, whereas the relation between Q_1 and coverage is more nuanced. That relation depends on T and on the interactions

between Q_1 and the rest of the $Q_k \geq 2$. Suppose that the frequency distribution of the sampled properties contains only singletons, i.e., $Q_k \geq 2 = 0$. From Eq. (11), that implies that coverage will be zero regardless of sample size (T). We will return to these issues again in the “The necessary simplifications” subsection.

Returning to Table 1, Eqs. (9) and (10) can be used to estimate the additional sampling effort necessary to achieve a certain desired coverage and the corresponding \hat{S} . Equation (9) can be used to estimate the number of extra participants for each concept (t^*), which would produce a similar coverage for each concept. For example, if we want to reach a coverage for each concept similar to the highest one already attained (78% for *gratitude*), Table 1 shows the corresponding additional number of participants for each concept (t^*) that we need. Note that t^* goes from as small as 2 (for *profit*) to as high as 88 (for *guilt*). Equation (10) allows estimating the expected

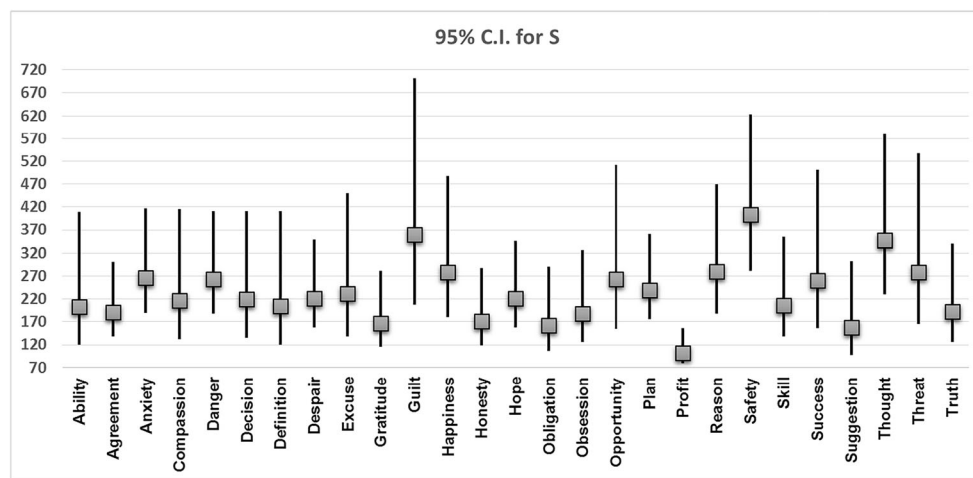


Fig. 2 Point estimates for semantic richness (\hat{S}) and corresponding 95% CI for each of the 27 concepts in CPN

semantic richness derived from adding t^* participants to the sample for each concept (last column in Table 1).

Given that the expressions applied in computing all the previously presented values might be at first somewhat difficult to understand and correctly use, we recommend to the interested reader recreating all the figures shown in Table 1 by introducing the values of Q_1 , Q_2 , T , S_{obs} and \mathcal{U} into the corresponding equations (3) to (10). Note that we used a rounded value of 78% for \hat{C} for *gratitude* (a more precise value is 0.77829). However, given the ceiling function, using Eq. (9) will produce a t^* for *gratitude* equal to 1 (in contrast to a value equal to 0 in Table 1). Nevertheless, we believe that approximations in Table 1 contribute to making it more readily understandable. Also note that when Eq. (9) gives a t^* above $2T$, then the actual t^* that must be used is $2T$, and that is the value to be inputted into Eqs. (8) and (10) to calculate $\hat{C}(T + t^*)$ and $\hat{S}(T + t^*)$, see footnote to Table 1.

Possible solutions to the threefold problem

In this section, we will integrate all the information presented in preceding sections, providing potential solutions to the three problems identified in the introductory section. To facilitate the discussion, we will focus first on when should CPNs be considered comparable and the problem of their replicability; secondly, we will focus on defining a heuristic method to determine sample size in CPN studies; and lastly, we will examine consequences for experimenters who use CPN data to select carefully controlled stimuli in terms of the reliability and generalizability of their results. On closing, we will discuss the necessary simplifications of the model in Eqs. (1) and (2).

The overall idea underlying our discussion is that, generally, researchers that collect CPN data are interested on the properties themselves (i.e., their identity) and their corresponding frequencies, more than on any single metric.

Because any measure currently computed from CPNs is derived from the properties' frequency distribution (e.g., different measures of semantic richness, different measures of similarity, property dominance, cue validity, property distinctiveness), appropriately characterizing those distributions should be a main goal of researchers.

Making concepts and CPNs comparable

As discussed in the Introduction section, there is currently no good way to compare results from CPNs. Trying to show that two CPNs are not different by betting on null results poses obvious problems. Here, we want to argue that comparisons across concepts and across CPNs can be meaningfully performed when coverages are sufficiently similar, and not necessarily when sample sizes are standardized (i.e., set to equal values across concepts), because concepts with the same coverage make sure that the not yet sampled properties constitute the same proportion of the total properties in the population (for the same argument made in ecology, see Chao & Jost, 2012 and Rasmussen & Starr, 1979). Adopting the decision rule of collecting data until a certain conventional estimated coverage ($\hat{C}(T)$) is achieved would allow for a better interpretation of differences between concepts within the same CPN, and also between CPNs.

We offer two examples of how comparisons across concepts could benefit from standardizing coverage instead of sample size. In the case of semantic richness research, as our analyses show and Table 1 illustrates, S_{obs} is influenced by sample size (i.e., increasing the number of participants, increases the probability of additional properties being produced). However, sample size operates in conjunction with the distribution of properties in the population. Thus, when sample sizes are standardized across concepts (i.e., all concepts sampled with the same number of participants), as is routinely done in CPNs, S_{obs} becomes only a rough estimator

of the true semantic richness (SR). This limits the precision with which we may compare different concepts along that same dimension. Take for example the data corresponding to *happiness* and *honesty* ($T = 37$), and to *definition* and *success* ($T = 30$) of our own CPN. Given that those two pairs of concepts have equal sample size, then based on S_{obs} one could state that *happiness*'s SR is larger than that of *honesty* ($92 > 74$), and that *definition*'s SR is smaller than that of *success* ($64 < 77$). However, Fig. 2 shows that the SR CIs for those pairs of concepts overlap, which suggests that those differences in SR are not statistically significant (for an $\alpha = 0.05$). Of course, this does not invalidate the use of S_{obs} as a semantic richness (SR) measure, but it highlights it being only an approximate measure. Regarding this last point about CIs, and to avoid confusion, note that variables such as SR and other similar measures are all random variables (e.g., given that there is not a single unique semantic structure in people's minds). Thus, collecting unbiased property frequency distributions and being able to compute CIs, as discussed in the current work, would allow estimating how precise the aforementioned approximations are, depending on CI width.

Something similar occurs with other metrics that can be computed from CPNs. Take for example a study by Wiemer-Hastings and Xu (2005), where concepts were compared in terms of the types of content of properties produced in a PLT (i.e., entity properties, relational properties, experiential content). According to the authors, their results showed that abstract concepts involved more experiential content, less entity properties and more relational properties than concrete concepts. Obviously, their conclusions depend on conceptual properties being sufficiently sampled (an unknown in their research). This becomes even more important with more fine-grained content types (e.g., dividing experiential content into emotional and cognitive; dividing relational properties into physical and social). Thus, we believe that sample size standardization sets limits to estimation precision for metrics that are computed from property frequency distributions.

Another advantage of adopting the decision rule of conducting CPNs until a certain conventional estimated coverage ($\hat{C}(T)$) is achieved, is that it would allow to meaningfully compare CPNs. To sensibly compare CPNs' results, one should ensure that those comparisons are not unduly affected by differences in the corresponding CPN studies' representativeness (i.e., sampling procedures, number of participants, etc.). Having collected and analyzed CPN data ourselves, we believe that researchers may be all too aware that even slight variations in procedures may lead to significant differences, making those differences difficult to interpret. For example, data collection procedures, cultural differences, and other similar variables could affect the number of properties participants can produce, something that would have an effect on property distributions, introducing perhaps spurious

differences in CPN data. In contrast, if sampling by standardizing $\hat{C}(T)$, researchers could ensure that concepts are sampled with the same degree of completeness, alleviating the aforementioned concerns by making comparisons possible regardless of sampling effort.

A heuristic for standardizing by coverage and determining sample size

Recall that, traditionally, sample sizes in CPN studies have been determined arbitrarily. Furthermore, the same number of participants is typically used for all concepts in a single CPN. Having used a different number of participants in our own CPN allowed us to explore the effect that sample size has on results. Though increasing sample size increases coverage, its effect is limited by the property frequency distribution of each particular concept. To illustrate by means of Table 1, consider the case of *happiness* and *honesty*. Both share the same sample size ($T = 37$), but note that the increase in sample size (t^*) necessary to achieve a 78% coverage is different for each (respectively, 36 and 16). The difference is explained precisely by differences in their property frequency distributions. *Happiness* has 65 singletons, versus the 49 singletons of *honesty*, meaning that, though T has the same value for both, the completeness of property sampling differs between them, presumably due to the shape of the property distribution in the population. From the preceding discussion, we can conclude that 37 participants allowed a better sampling of *honesty* than of *happiness*. Thus, as already discussed, sample size in CPNs does not necessarily have to be the same across all concepts. In fact, the same sample size for all concepts will probably lead to unfair comparisons.

How then should researchers decide on appropriate sample sizes? We propose here that Eqs. (3) through (10) provide researchers an informed means of deciding which concepts she might select to apply the additional effort to match their coverages (a more in depth discussion of coverage standardization can be found in Chao & Just, 2012, and Rasmussen & Starr, 1979). In a nutshell, we propose that researchers use a two-stage sampling procedure. In the first stage, researchers should conduct a PLT for each concept with a small number of participants (judging from the literature, ten or perhaps 15 participants per concept could suffice). With those data, it would be possible to use Eq. (7) to estimate the current estimated coverage, and Eq. (9) to estimate the t^* additional participants necessary for achieving a desired coverage \hat{C}_{target} . For example, if with our own CPN we wanted to reach a coverage for each concept similar to the highest coverage already attained (78% for *gratitude*), Table 1 shows the corresponding additional number of participants for each concept (t^*) that we need. Note that t^* goes from as small as 2 (for

profit) to as high as 88 (for *guilt*). Thus, the researcher now has an informed means of deciding if she might apply the additional effort to match the coverages for all concepts, or just for some of them; perhaps the ones that are theoretically more interesting.

Note that a more general strategy would be to use true incremental sampling (i.e., increasing sample size one participant at a time). Beginning with a small sample of participants, maybe 10 or 15, sample size could be increased by one participant at a time until the desired coverage was reached. Though we believe this would be an attractive strategy, given that it would tend to optimize sampling effort, we also believe that meeting the necessary conditions would be difficult. Increasing sample size one participant at a time, assumes we could compute coverage dynamically and in real time, and to decide when to stop collecting accordingly. Because in typical CPN studies, whole phrases need to be coded into property types prior to any analysis, dynamic coverage computations would be costly (i.e., incremental sampling would entail solving the problem of how to perform whole phrase incremental coding). A different coding procedure, such as the bag-of-words approach, which is amenable to automating (Buchanan, De Deyne, & Montefinese, *in press*), could allow true incremental sampling, but discussing this alternative is well beyond the scope of the current work.

In the above analysis, one should also take into account that, given the values of the variables involved in the estimation of the coverage for concepts, the expected additional coverage attained from the extra number of participants varies among concepts. To illustrate this point, Fig. 3 shows how estimated coverage changes when increasing the number of participants for three different concepts.

In Fig. 3, note that increasing participants for *profit* pays off much more than for *compassion* and *thought* (no pun intended). Figure 3 shows that an increase of 50 participants increases coverage for *profit* in about 16.5%, whereas for *compassion* that increase is only 13.3% and for *thought* it is only 10.2%. Additionally, note that trying to reach a 100% coverage entails using a prohibitively large number of participants, especially for *compassion* and *thought*.

In addition to aiming for similar coverages, a researcher would also want to have reliable semantic richness (SR) estimates (\hat{S}). For example, in our CPN, note in Fig. 2 that for sample sizes used, S shows large CIs, suggesting that with the current sample sizes there are no significant differences between most concepts' semantic richness. In contrast, a good example of what a researcher should desire is given by the comparison of concepts *plan* and *profit*, which show small CIs (and in this case, also non-overlapping) and similar coverages (72% and 77%, respectively). Thus, the ideal case would be to jointly assess how many more participants per concept are needed to attain a certain coverage (t^* , per Eq. (9)), and also the CI width of the corresponding semantic

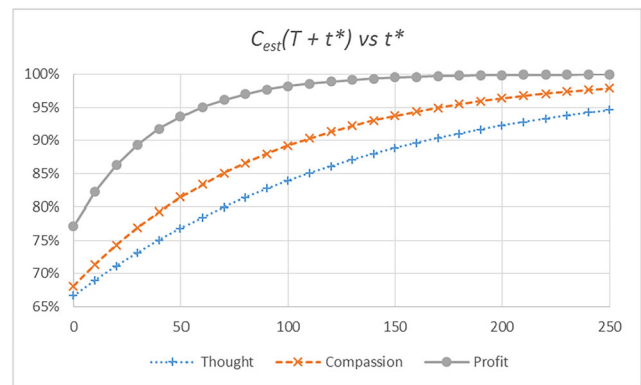


Fig. 3 Estimated coverage vs. extra number of participants for three different concepts

richness estimator $\hat{S}(T + t^*)$. However, if one peruses expressions (3) to (10), one can see that no equation exists to estimate the variance of $\hat{S}(T + t^*)$, and hence one cannot easily compute the corresponding CI. Consequently, there is no sure and easy manner to handle this, forcing the researcher to use his/her better judgment. Two elements that are available for that judgment are the current width of the S estimate's CI and the value of the estimated SR when adding t^* participants, i.e., $\hat{S}(T + t^*)$.

First, if for a given concept, the current coverage is high (not too far removed from the desired coverage), and the current CI for S is conveniently narrow, then the researcher can be reasonably confident that increasing sample size would pay off as expected (i.e., that the result of the additional sampling effort would be an adequate coverage with a reliable \hat{S} estimate). In our CPN, the concept *profit* serves as an example, i.e., it currently exhibits a 77% coverage and has a relatively narrow CI for S (78.9 to 156.9). Here we must clarify that a conveniently narrow CI is one that does not overlap with other concepts' CI for the parameters of interest, in our case specifically for \hat{S} CIs. As already mentioned, that assumes that researchers want to attain statistically significant differences among concepts' estimates of the parameters of interest.

Second, assuming that a researcher wants to find statistically significant differences, he/she can compare the estimated new coverages $\hat{S}(T + t^*)$ between the concepts that he/she wants to contrast in terms of SR, and approximately see whether those SR estimators are sufficiently different. If in fact the values for $\hat{S}(T + t^*)$ are different enough, then the additional sampling effort might be useful. Of course, given that one does not have an estimate of the corresponding CIs, a judgment call is needed. For example, in our CPN, if the researcher is interested in comparing the SR of *profit*, *thought* and *reason*, he/she might judge that the comparison between *profit* and *thought*, and between *profit* and *reason* might be informative, given that the corresponding $\hat{S}(T + t^*)$ are 64.0, 190.7, and 170.6, respectively. On the other hand, the comparison between *thought* and *reason* might prove to be useless.

Consequences for users of CPNs

In previous work (Chaigneau, Canessa, Barra, & Lagos, 2018), we have argued that weeding out low-frequency properties reduces data variability. Because much information in CPNs is carried by variability, rather than cleaning noise from data, weeding out may in actuality be throwing away relevant information. In the same vein, the current work highlights that low-frequency properties are important because they contain information about the likelihood of yet-to-be-sampled properties. As we have shown above, adopting the otherwise reasonable criterion of standardizing sample sizes, results in imprecise estimations of semantic richness, as measured by S_{obs} . Thus, results based directly on those measures are bound to be rough estimations that are likely to reveal only broad patterns or strong effects in the data. Importantly, note that we are not holding that prior research with CPNs is wrong or is not useful. Our own CPN attests to their usefulness. Rather, we are holding that because those results are by necessity only rough approximations, they reveal perhaps stable, but nevertheless only broad patterns, and that their generalizability is limited to the particular CPN under consideration. Furthermore, because other measures that can be computed from CPN data (e.g., similarity measures, other semantic richness estimates) will be affected in similar ways by sampling size, sampling quality, and by the particular details of property frequency distributions, results from studies that use those metrics computed from the current CPNs are also limited in generality.

A critical reader may object that collecting data from very large samples should appease our concerns (e.g., De Deyne et al., 2019). There are several things to note regarding this objection. First, though De Deyne et al.'s study reports 100 participants per each of their over 12,000 cue words (i.e., they standardize sample size), achieving that sample size required more than 80,000 participants over a 7-year period. That reflects an enormous effort of data collection and data handling, which would be even greater for a CPN study. Note that De Deyne et al. collected association data, which is less cumbersome to handle than CPN data. In that study, participants were asked to produce three associates (single words, not sentences) per cue. In contrast, in CPN studies, participants will typically list well over the three-property limit, and those properties are typically sentences, not single words, all of which makes very large samples less feasible for CPNs because of the necessary additional complications of data cleaning and coding. Furthermore, and precisely because of the enormous effort necessary for collecting data such as those of De Deyne et al., we cannot help but wonder whether standardizing coverage might not have been a more practical strategy. It is likely that such large studies spanning many years are more likely to introduce error due to many different factors that characterize census-type and longitudinal studies, and

standardizing coverage could have been used to reduce sampling effort, as previously discussed.

The necessary simplifications

As with any model, Eqs. (1) and (2), which are the basis for deriving the rest of the expressions, make some simplifications. Here we discuss those simplifications and argue that they are reasonable for modeling PLT and CPN data. We note that many of our arguments are similar to the ones used and deemed acceptable in justifying the application of the model in ecology (Chao & Chiu, 2016 and references therein).

A first assumption is that the detection probability for each property i (π_i) is independent from the rest. Though this is a generally accepted simplification (e.g., computing property frequencies and using cosine similarity immediately assumes independence, e.g., McRae, Cree, Seidenberg & Mcnorgan, 2005), it is probably not completely true at the level of each individual participant (i.e., properties might be correlated, such that evoking a given property affects the probability of evoking the following one). In what follows, we discuss in some detail why independence is a reasonable assumption for CPNs. In particular, we discuss why inter-feature correlation might not be problematic.

In experimental studies, inter-feature correlations have for quite some time been thought to be of theoretical interest (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Many studies have found that people can become sensitive to inter-feature correlations, in particular when they are required to predict the value of unknown features based on the values of other known features (e.g., Anderson & Finchman, 1996). We believe similar results have been replicated many times since then.

If inter-feature correlations were prevalent in CPN data, then the independence assumption could be questioned. In CPNs, there are at least two different ways to conceptualize inter-feature correlations. First, it may be that two properties (A and B) are more probable conditional on a given concept (C) and less probable given a different concept (it is possible that A and B, e.g., “barks” and “wags its tail”, are more likely to occur together conditional on the category *dog*). In other words, two properties A and B are correlated if $p(A \wedge B|C) > p(A \wedge B|\neg C)$. This is undoubtedly true in concepts (e.g., no other concept other than *dog* will produce “barks” and “wags its tail”). Note, however, that this type of correlation is not problematic for our assumption because our calculations assume independence of properties within concepts, and not between concepts.

The other way to conceptualize inter-feature correlations is the idea of “chaining”. In chaining, the probability of feature B changes as a function of whether feature A was produced prior to it or not ($p(B|A) > p(B|\neg A)$). This could be problematic because, if properties come in packages, then counting the

number of properties by the number of coded properties would overestimate S_{obs} . However, evidence suggests that chaining in CPNs might not be as problematic as experimental studies would suggest. De Deyne, Navarro, Perfors, Brysbaert, and Storms (2019) looked for evidence of chaining in C-A-B triplets (i.e., cueing concept C, property A, property B), finding strong evidence of chaining in 1% of the triplets, and moderate evidence in only 19% of the cases. These results led the authors to conclude that only a modest amount of response chaining (i.e., dependence) existed in their data. In other words, the amount of correlation or chaining for a given A-B pair of properties in a typical CPN may be a random variable exhibiting only a modest correlation (i.e., a positive correlation in some individuals, but a negative one in others, such that those correlations tend to cancel each other out across individuals). Another alternative worth considering, which also speaks in favor of the independence assumption, is that it is possible that people, though actually producing response chains, produce different chains of responses (e.g., someone lists “turbine” right after “wing” when listing properties for *airplane*, whereas someone else lists “propeller” right after “wing”). Because CPNs accumulate data across many individuals, this would turn properties independent across individuals.

To close our discussion on independence, even if chaining existed and was highly prevalent in CPNs, please note that the independence assumption could be relaxed if dependencies were possible to be estimated. In such a case, one would have to resort to numerical methods (e.g., bootstrapping), which are already used in ecological research and may be readily applied to CPN data (Chao & Chiu, 2016; Chao, Gotelli, Hsieh, Sander, Ma, Colwell & Ellison, 2014). We refrain from discussing this topic further, and defer it to future work.

A second related simplification is that there are no participant-specific effects. In the case of the PLT, this means that each participant is equally representative of the population in generating properties, i.e., that there are no systematic and important differences among participants related to the listing process. As already discussed above, that is a reasonable assumption, and in fact one that is routinely done in CPN studies, if using a random selection of participants. Note, just as occurs for the independence assumption, that there are models which can quite simply accommodate those participant-specific effects (Chao & Chiu, 2016), but the problem lies in determining and quantifying those effects.

A third simplification is that the model assumes that a concept is described by a finite number of properties. We believe this is a reasonable assumption, not because we assume that the underlying un verbalized semantic properties are finite, but because property production is confined to a very specific moment in time (i.e., the time span in which the PLT is carried out). Even if the different factors that limit property production (e.g., cognitive load, interference) were somehow

suspended, the total number of properties accessible to participants for report at any given moment is likely to be finite, though perhaps very large. If, on the contrary, property lists were collected during a long period of time, then many factors could make the total list of properties increase indefinitely in length (e.g., cultural change, conceptual drift), which would be undesirable if the goal were to obtain information about the structure of semantic memory.

A fourth and final simplification is assuming that each listed token can be coded into one and only one property type. In incidence matrices, any given response is a token of only one property type, depending on decisions made during coding. However, just as exemplars can belong to multiple categories (e.g., a cat is a feline and also a pet), tokens of conceptual properties could be classified in multiple property types (e.g., the token “achieving a goal” could be coded as “goal” as well as “achievement”). Choosing a single code for each token property is related to the coding process and its reliability. The current and accepted practice to deal with that problem is to have more than one coder, who independently code the token responses, and then assess inter-coder agreement. Although researchers know that that form of dealing with the coding process is not ideal, at present it is the best we have. Hence, this fourth simplification is not a new one and also applies to all CPN studies, which means that applying it here does not unduly invalidate our model.

A case study: Revisiting our CPN

As suggested by an anonymous reviewer, and to illustrate the relevance of coverage for designing experiments and obtaining valid results, aside from its relevance for estimating S_{obs} , in the present case study, we use our own CPN data to search for evidence of a relation between a concept’s mean list length (i.e., the mean number of properties produced by participants for a given concept) and its associated properties’ mean dominance (i.e., the mean frequency of those properties that are produced in response to the cueing concept). In general, CPN data should show that, as concepts’ mean list length increases, concepts’ mean property dominance decreases. As will become clear next, there is evidence suggesting that we should replicate this relation in our norms. But, more importantly, if coverage were irrelevant for practical purposes, then we should find that it does not affect the probability of confirming the hypothesized relation (or that its effect can be easily explained away).

Supporting literature

Many variables may affect a property’s dominance. Depending on the task at hand, e.g., property generality (i.e., the percentage of concepts in a CPN for which a certain

property is produced) or property distinctiveness (i.e., $100 - \text{property generality}$) may dictate which properties become relevant and dominate participants' cognitive processing (Devereux, Taylor, Randall, Geertzen, & Tyler, 2016; Duarte, Marquié, Marquié, Terrier, & Ousset, 2009; Flanagan, Copland, Chenery, Byrne, & Angwin, 2013; Grondin, Lupker, & McRae, 2009; Siew, *in press*). Whatever the reason for a property's dominance, the literature suggests that those properties that are more accessible or dominant (i.e., those that are found frequently in participants' lists) are also those that tend to be produced first (i.e., those that are likely to become available sooner) (Montefinese, Ambrosini, Fairfield, & Mammarella, 2013; Ruts, De Deyne, Ameel, Vanpaemel, Verbeemen, & Storms, 2004). This relation directly implies that the longer a list is, the lower will be the list's mean dominance. In fact, we have found this relation in four different CPNs across four different countries and three different languages (Canessa & Chaigneau, *in press*; Chaigneau, Canessa, Barra, & Lagos, 2018). Furthermore, we found that the specific functional form of the relation is hyperbolic (i.e., $d = b_0 + \frac{b_1}{s}$, where $d = \text{dominance}$, $s = \text{mean list length}$, b_0 and $b_1 = \text{coefficients to be estimated from the data}$).

Empirical analysis supporting the effects of coverage on results

Recall that our goal in this section is to show that coverage is relevant for being able to detect relations between variables based on CPN data, and that though evidently any study or experiment provides estimations of those relations, what is at stake here is the quality of those estimations in terms of the true population parameters. Recall also that our CPN data include concepts showing different coverages, due to the interaction between sample size and the characteristics of each concept's property frequency distribution. This feature of our CPN study allowed us to mimic two separate studies, one with lower coverage data (as may happen if sample size is standardized without regards to property frequency distribution), and one with higher coverage data.

To mimic the two aforementioned studies, we divided our CPN's 27 concepts by their calculated coverage \hat{C} , thus producing a lower coverage (less than or equal to 67%) and a higher coverage (more than or equal to 68%) group of concepts (see Table 1). The respective mean coverages for each group are 61.0% and 71.1%, which are significantly different ($t(25) = 6.602$, $p < .001$). Producing two groups with a similar number of concepts in each (13 concepts for the low coverage group, and 14 concepts for the higher coverage group), allowed us avoiding arbitrary decisions regarding how to select concepts for each group (e.g., we did not resort only to concepts with extremely high or extremely low coverage values).

Assume now that each group represents a different study with the goal of testing the inverse relation between mean dominance and mean list length across concepts. To this end, for each group of concepts, values needed for the hyperbolic equation were computed (i.e., d and s), and the b_0 and b_1 coefficients were estimated by using ordinary least squares (OLS). Results are presented in Fig. 4.

As Fig. 4a shows, doing the study with the lower coverage concepts produced a curve that runs counter to prior evidence. Although the regression equation exhibits the hyperbolic form, it suggests that average dominance increases with average list length ($d = 12.141 - 35.186 / s$, $R^2 = 0.503$, $F(1,11) = 11.124$, $p = .007$). Note that this study would have concluded that concepts for which people list a large number of properties are also concepts with overall high dominance properties, something that cannot be reconciled with prior literature (Montefinese, Ambrosini, Fairfield, & Mammarella, 2013; Ruts, De Deyne, Ameel, Vanpaemel, Verbeemen, & Storms, 2004; Canessa & Chaigneau, *in press*; Chaigneau, Canessa, Barra, & Lagos, 2018). In stark contrast, Fig. 4b shows that doing the same study with the higher coverage concepts produced a curve that replicates previous findings in the literature. The OLS procedure yields a significant hyperbolic regression that inversely relates d and s ($d = -1.326 + 39.679 / s$, $R^2 = 0.286$, $F(1,12) = 4.798$, $p = .049$). This result is evidently consistent with prior literature.

Results from our case study suggest several conclusions. First, coverage matters, and not taking it into account may lead to erroneous and misleading results. Note that although the difference in coverage between both groups is rather small (10.1%), it still has an important impact on results. Second, though it is possible that some of the problems of low coverage could be solved by simple brute force (i.e., adding participants or adding concepts), using the strategy of sampling to standardize coverage is probably a better way to achieve a reasonable trade-off between sampling effort and estimation accuracy. Finally, because, as we have shown, the relation of sample size and coverage is contingent on the distributional characteristics of conceptual properties (i.e., $Q_1, Q_2, Q_3, \dots, Q_T$), increasing the number of participants and standardizing sample size may still introduce distortions on many estimated values, such that comparisons between concepts and across CPNs become problematic. This is yet another argument against a brute force approach for solving the problems that our case study makes evident.

Conclusions

At present, CPN studies and corresponding PLTs can deliver data from which one can calculate only rough estimators of the true unknown population parameters of interest. Also, given that there are no means of assessing the

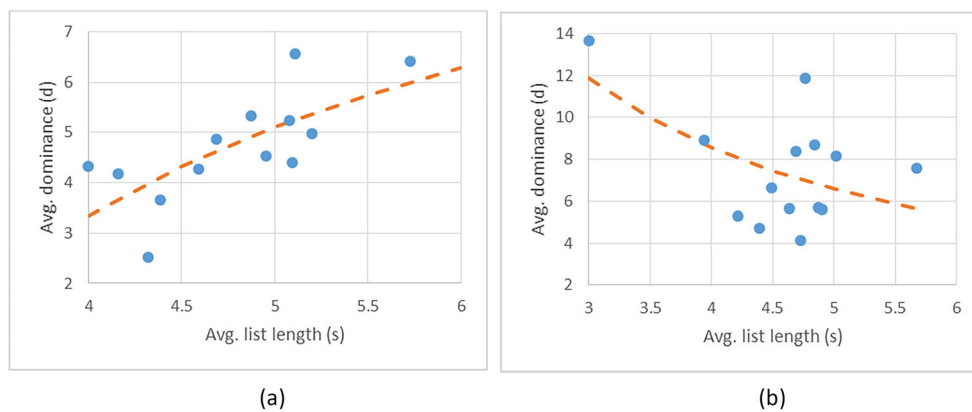


Fig. 4 Average dominance (d) of concepts vs. average list length (s). Panel **a** is for concepts with the lowest coverage and **b** for concepts with highest coverage. Each concept is represented by *dots*, and the non-linear regression equations are represented by *dotted lines*

reliability of the estimated parameters (i.e., their variability), it is a common practice to treat those point estimates as true population values. Given these state of affairs in CPN studies, only broad results can be obtained from analyzing those parameters, as well as when assessing the relation among them and to other variables obtained in experiments. Associated to those issues is the problem of determining sample size for a PLT (i.e., the number of participants who will list properties for each concept). The traditional consensus is to use between 20 and 30 participants for each concept and standardize that number across concepts, intuitively believing that using the same number of participants for each concept will render the estimated parameters comparable across concepts. Also, the general belief is that the more participants a PLT has, ipso facto, the better the parameter estimates will be. Contrarily, in the current work, we suggest viewing the PLT as a parameter estimation procedure, where we obtain only estimates of the true unknown population parameters. Thus, more meaningful and fine-grained analyses of those parameters must consider the variability of their estimators. To that end, we introduce a model from the field of ecology, which can be applied to calculate some of the parameters of a PLT and their corresponding variances. Additionally, the expressions derived from that model can be used to guide the sampling process, and particularly to estimate a sensible number of participants for each concept and to assess whether that number is feasible. We argue that the number of participants must not necessarily be the same for each concept, but that it should be determined so that concepts' coverages are approximately the same. That may allow more reasonable comparisons of parameter values among different concepts within a CPN, and between different CPNs. As an illustration of the practical relevance of concepts' property coverage, we used CPN data collected in our laboratory to mimic a lower- and a higher-coverage study's ability to replicate an empirical association obtained in prior research.

One limitation of the current work is that the presented formulae, although easily applicable because of their closed mathematical form, do not allow calculating estimators and their standard errors for all the potential parameters derived from CPN data. Thus, at present, the application of the expressions is limited to a few parameters. Nevertheless, we firmly believe that this work is a first important step in the right direction. It not only allows estimating some important parameters and sensibly establishing sample sizes, but in a broader sense, it exposes the need of advancing in treating CPN studies as parameter estimation research. Hence, we think that devoting research to expanding the application of the presented ideas is worth the effort. In doing so, and as a means of dealing with the difficulties in advancing the mathematical model, we think that bootstrapping methods can be used to calculate estimates and their standard errors for other interesting parameters. In fact, as research in ecology shows (Chao & Chiu, 2016; Chao, Gotelli, Hsieh, Sander, Ma, Colwell, & Ellison, 2014), bootstrapping methods allow calculating point estimators and their standard errors based on collected data without having to mathematically specify the underlying models. We believe that similar procedures may fruitfully be applied to PLT and CPN data. Our current work, which imports methods and models from ecology into CPN research, attests to that fact.

Acknowledgements This research was carried out with funds provided by ANID, Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) of the Chilean government grant 1200139. Felipe Medina thankfully acknowledges funding from Comisión Nacional de Investigación Científica y Tecnológica, CONICYT Ph.D. fellowship 21151523. We are grateful to Penny Pexman and Jorge Vivas for their valuable input on the current work. The authors declare to have no known conflicts of interest regarding the work being reported here.

None of the data or materials for the experiments reported here are available, and none of the experiments was preregistered. However, the interested reader may directly contact the corresponding author and he will send the corresponding materials and data to the reader.

References

- Anderson, J. R., & Fincham, M. (1996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *22*, 259–277. <https://doi.org/10.1037/0278-7393.22.2.259>
- Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233. doi:<https://doi.org/10.1006/jmps.1995.1021>
- Ashcraft, M. H. (1978). Property norms for typical and atypical items from 17 categories: A description and discussion. *Memory & Cognition*, *6*(3), 227–232. doi:<https://doi.org/10.3758/BF03197450>
- Barsalou, L.W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101–140). Cambridge: Cambridge University Press.
- Bolognesi, M., Pilgram, R., & van den Heerik, R. (2017). Reliability in content analysis: The case of semantic feature norms classification. *Behavior Research Methods*, *49*(6), 1984–2001. doi:<https://doi.org/10.3758/s13428-016-0838-6>
- Bruffaerts, R., De Deyne, S., Meersmans, K., Liuzzi, A. G., Storms, G., & Vandenberghe, R. (2019). Redefining the resolution of semantic knowledge in the brain: Advances made by the introduction of models of semantics in neuroimaging. *Neuroscience and Biobehavioral Reviews*, *103*, 3–13. doi:<https://doi.org/10.1016/j.neubiorev.2019.05.015>
- Buchanan, E.M., De Deyne, S., & Montefinese, M. (in press). A practical primer on processing semantic property norm data. *Cognitive Processing* doi: <https://doi.org/10.1007/s10339-019-00939-6>
- Canessa, E. & Chaigneau, S. E. (in press). Mathematical regularities of data from the property listing task. *Journal of Mathematical Psychology*, 102376. doi:<https://doi.org/10.1016/j.jmp.2020.102376>
- Chaigneau, S. E., Canessa, E., Barra, C., & Lagos, R. (2018). The role of variability in the property listing task. *Behavior Research Methods*, *50*(3), 972–988. doi:<https://doi.org/10.3758/s13428-017-0920-8>
- Chao, A., & Chiu, C. H. (2016). Species richness: Estimation and comparison. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri & J. L. Teugels (Eds.) *Wiley StatsRef: Statistics Reference Online* (pp. 1–26). Chichester, UK: John Wiley & Sons, Ltd doi: <https://doi.org/10.1002/9781118445112.stat03432.pub2>.
- Chao, A. & Jost, L. (2012). Coverage-based rarefaction: standardizing samples by completeness rather than by sample size. *Ecology*, *93*, 2533–2547.
- Chao, A., Gotelli, N., Hsieh, T.C., Sander, E., Ma, K.H., Colwell, R. & Ellison, A. (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs*, *84*(1), 45–67.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi: <https://doi.org/10.1177/001316446002000104>
- Coley, J. D., Hayes, B., Lawson, C., & Moloney, M. (2004). Knowledge, expectations, and inductive reasoning within conceptual hierarchies. *Cognition*, *90*(3), 217–253. doi:[https://doi.org/10.1016/S0010-0277\(03\)00159-8](https://doi.org/10.1016/S0010-0277(03)00159-8)
- Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, *132*(2), 163–201. doi:<https://doi.org/10.1037/0096-3445.132.2.163>
- De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The “Small world of words” English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006. doi:<https://doi.org/10.3758/s13428-018-1115-7>
- Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The centre for speech, language and the brain (CSLB) concept property norms. *Behavior Research Methods*, *46*(4), 1119–1127. doi:<https://doi.org/10.3758/s13428-013-0420-4>
- Devereux, B. J., Taylor, K. I., Randall, B., Geertzen, J., & Tyler, L. K. (2016). Feature statistics modulate the activation of meaning during spoken word processing. *Cognitive Science*, *40*(2), 325–350. doi: <https://doi.org/10.1111/cogs.12234>
- Duarte, L. R., Marquié, L., Marquié, J., Terrier, P., & Ousset, P. (2009). Analyzing feature distinctiveness in the processing of living and non-living concepts in Alzheimer's disease. *Brain and Cognition*, *71*(2), 108–117. doi:<https://doi.org/10.1016/j.bandc.2009.04.007>
- Flanagan, K. J., Copland, D. A., Chenery, H. J., Byrne, G. J., & Angwin, A. J. (2013). Alzheimer's disease is associated with distinctive semantic feature loss. *Neuropsychologia*, *51*(10), 2016–2025. doi: <https://doi.org/10.1016/j.neuropsychologia.2013.06.008>
- Garrard, P., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, *18*(2), 125–174. doi:<https://doi.org/10.1080/02643290125857>
- Glaser, B.G. & Strauss, A.L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.
- Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200. doi:<https://doi.org/10.1037/0096-3445.123.2.178>
- Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In M. Oaksford and N. Chater (Eds.). *The probabilistic mind: Prospects for rational models of cognition*. Oxford: Oxford University Press.
- Grondin, R., Lupker, S. J., & McRae, K. (2009). Shared features dominate semantic richness effects for concrete concepts. *Journal of Memory and Language*, *60*(1), 1–19. doi:<https://doi.org/10.1016/j.jml.2008.09.001>
- Hallgren K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. doi:<https://doi.org/10.20982/tqmp.08.1.p023>
- Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *18*(4), 441–461. doi:[https://doi.org/10.1016/S0022-5371\(79\)90246-9](https://doi.org/10.1016/S0022-5371(79)90246-9)
- 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. doi:<https://doi.org/10.1016/j.cognition.2014.01.001>
- Hough, G., & Ferraris, D. (2010). Free listing: A method to gain initial insight of a food category. *Food Quality and Preference*, *21*(3), 295–301. doi:<https://doi.org/10.1016/j.foodqual.2009.04.001>
- Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelle*, *37*, 241–272.
- Kounios, J., Green, D. L., Payne, L., Fleck, J. I., Grondin, R., & McRae, K. (2009). Semantic richness and the activation of concepts in semantic memory: Evidence from event-related potentials. *Brain Research*, *1282*, 95–102. doi:<https://doi.org/10.1016/j.brainres.2009.05.092>
- Kremer, G., & Baroni, M. (2011). A set of semantic norms for German and Italian. *Behavior Research Methods*, *43*(1), 97–109. doi:<https://doi.org/10.3758/s13428-010-0028-x>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. doi:<https://doi.org/10.2307/2529310>
- Lenci, A., Baroni, M., Cazzolli, G., & Marotta, G. (2013). BLIND: A set of semantic feature norms from the congenitally blind. *Behavior Research Methods*, *45*(4), 1218–1233. doi:<https://doi.org/10.3758/s13428-013-0323-4>

- McRae, K., Cree, G. S., Westmacott, R., & De Sa, V. R. (1999). Further evidence for feature correlations in semantic memory. *Canadian Journal of Experimental Psychology*, 53(4), 360–373. doi:<https://doi.org/10.1037/h0087323>
- McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4), 547–559. doi:<https://doi.org/10.3758/BF03192726>
- Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013). Semantic memory: A feature-based analysis and new norms for Italian. *Behavior Research Methods*, 45(2), 440–461. doi:<https://doi.org/10.3758/s13428-012-0263-4>
- Perri, R., Zannino, G., Caltagirone, C., & Carlesimo, G. A. (2012). Alzheimer's disease and semantic deficits: A feature-listing study. *Neuropsychology*, 26(5), 652–663. doi:<https://doi.org/10.1037/a0029302>
- Pexman, P. M., Hargreaves, I. S., Edwards, J. D., Henry, L. C., & Goodyear, B. G. (2007). The neural consequences of semantic richness: When more comes to mind, less activation is observed: Research report. *Psychological Science*, 18(5), 401–406. doi:<https://doi.org/10.1111/j.1467-9280.2007.01913.x>
- Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin and Review*, 15(1), 161–167. doi:<https://doi.org/10.3758/PBR.15.1.161>
- Rasmussen, S.L., & Starr, N. (1979). Optimal and adaptive stopping in the search for new species. *Journal of the American Statistical Association*, 74, 661–667.
- Recchia, G., & Jones, M. N. (2012). The semantic richness of abstract concepts. *Frontiers in Human Neuroscience*, 6, 315. doi:<https://doi.org/10.3389/fnhum.2012.00315>
- Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4), 573–605. doi:[https://doi.org/10.1016/0010-0285\(75\)90024-9](https://doi.org/10.1016/0010-0285(75)90024-9)
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3), 382–439. [https://doi.org/10.1016/0010-0285\(76\)90013-X](https://doi.org/10.1016/0010-0285(76)90013-X)
- Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers*, 36, 506–515. doi:<https://doi.org/10.3758/BF03195597>
- Schyns, P. G., Goldstone, R. L., & Thibaut, J. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, 21(1), 1–54. doi:<https://doi.org/10.1017/S0140525X98000107>
- Siew, C. S. Q. (in press). Feature distinctiveness effects in language acquisition and lexical processing: Insights from megastudies. *Cognitive Processing* doi:<https://doi.org/10.1007/s10339-019-00947-6>
- Sokal, R., & Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409–1438.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327–352. doi:<https://doi.org/10.1037/0033-295X.84.4.327>
- Tversky, A., & Hutchinson, J. W. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, 93(1), 3–22. doi:<https://doi.org/10.1037/0033-295X.93.1.3>
- Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, 48(4), 422–488. doi:<https://doi.org/10.1016/j.cogpsych.2003.09.001>
- Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1), 183–190. doi:<https://doi.org/10.3758/BRM.40.1.183>
- Vivas, J., Vivas, L., Comesaña, A., Coni, A. G., & Vorano, A. (2017). Spanish semantic feature production norms for 400 concrete concepts. *Behavior Research Methods*, 49(3), 1095–1106. doi:<https://doi.org/10.3758/s13428-016-0777-2>
- Walker, L. J., & Hennig, K. H. (2004). Differing conceptions of moral exemplarity: Just, brave, and caring. *Journal of Personality and Social Psychology*, 86(4), 629–647. doi:<https://doi.org/10.1037/0022-3514.86.4.629>
- Wiemer-Hastings, K., & Xu, X. (2005). Content Differences for Abstract and Concrete Concepts. *Cognitive Science*, 29, 719–736. doi:https://doi.org/10.1207/s15516709cog0000_33
- Wu, L., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, 132(2), 173–189. doi:<https://doi.org/10.1016/j.actpsy.2009.02.002>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.