CrossMark

Which variables should be controlled for to investigate picture naming in adults? A Bayesian meta-analysis

Cyril Perret[1] · Patrick Bonin[2]

Abstract

Selecting items for designing psycholinguistic experiments can be a very hard and time-consuming process, because of the large number of variables that need to be controlled for. This is clearly the case for picture-naming experiments because, thanks to the collection of psycholinguistic norms on both pictures and their names, a large number of factors that affect naming speed and/or accuracy have been found. In the present study, a Bayesian meta-analysis was performed to determine the extent to which the variables that have generally been considered by researchers as important to control for are indeed worth taking into account. The meta-analysis revealed that most of the variables that are considered in picture-naming studies have a strong or very strong influence on naming speed (image agreement, name agreement, image variability/imageability, age of acquisition, and conceptual familiarity), whereas two variables that are very often taken into account (visual complexity and length) yielded null effects. The results were inconclusive for lexical frequency. At a methodological level, Bayesian meta-analyses constitute a very useful tool for guiding researchers when selecting materials for experiments.

Keywords Bayesian meta-analysis · Picture naming · Psycholinguistic norms

Ever since the days of Cattell (1886), researchers have worked hard to understand the cognitive processes and representations underpinning conceptually driven word production. To achieve this aim, they have mostly relied on picture naming. In this task, participants have to say aloud or write down as quickly as possible, while maintaining a level of accuracy, the names corresponding to pictured objects. Objects are often represented as black-and-white or colored line drawings, or as photographs. Two main reasons have been put forward to justify the use of this experimental task. First of all, object naming is a fast, efficient, and relatively effortless cognitive skill, especially in adults, and picture naming makes it possible to track in real time the processes and representations

underpinning object naming. Second, it is generally assumed that picture naming operationalizes a more natural communication situation in which speakers (or writers) wish to express an idea verbally. Thus, one assumption is that picture naming mobilizes the very same processes that are involved in conceptually driven word production (Bock & Levelt, 1994). Most views of word production (e.g., Bonin, Méot, Lagarrigue, & Roux, 2015; Glaser, 1992; C. J. Johnson, Paivio, & Clark, 1996; Levelt, Roelofs, & Meyer, 1999) assume that at least five levels of processing are involved in object naming: (1) perceptual analysis of the visual input, which results in activation of stored structural knowledge about the object; (2) a level corresponding to the retrieval of semantic/conceptual information; (3) lexical selection, which makes syntactic information such as grammatical category available—that is, the lemma level; (4) lexeme encoding (or name retrieval), which makes segmental and metrical information available; and (5) motor programming and execution.

Because picture naming requires the use of pictures, and since these can vary greatly in their individual characteristics, researchers have collected norms on sets of pictures and their names in order to make standardized stimuli available to the research community. The use of standardized pictures has permitted more reliable comparisons among the studies on picture

---

✉ Cyril Perret
    cyril.perret@univ-poitiers.fr

1   Université de Poitiers, Université de Tours, CNRS, 5, rue Théodore Lefebvre, TSA 21103, 86073 Poitiers Cedex 9, France

2   Université de Bourgogne, CNRS, Franche-Comté, France

Springer

naming. In 1980, Snodgrass and Vanderwart (SV) were the first to collect norms in American English for a large set of 260 black-and-white line drawings. Because there are cultural and linguistic variations in the ways the same pictures are named (Boukadi, Zouaidi, & Wilson, 2016; Sanfeliu & Fernandez, 1996), norms have been collected for the original SV pictures in several languages: American English (Snodgrass & Yuditsky, 1996), British English (Ellis & Morrison, 1998; Johnston, Dent, Humphreys, & Barry, 2010), Dutch (Severens, Van Lommel, Ratinckx, & Hartsuiker, 2005), French (Alario et al., 2004; Bonin, Chalard, Méot, & Fayol, 2002; Bonin, Peereman, Malardier, Méot, & Chalard, 2003; Perret & Laganaro, 2013; Valente, Bürki, & Laganaro, 2014), Italian (Dell'Acqua, Lotto, & Job, 2000), Icelandic (Pind & Tryggvadottir, 2002), Japanese (Nishimoto, Miyawaki, Ueda, Une, & Takahashi, 2005), Mandarin Chinese (Liu, Hao, Li, & Shu, 2011; Weekes, Shu, Hao, Liu, & Tan, 2007), Spanish (Cuetos, Ellis, & Alvarez, 1999), and Welsh (Barry, Morrison, & Ellis, 1997). In addition, normative studies have made possible the investigation of the factors influencing object-naming speed—that is, the duration between the onset of picture presentation and the beginning of the participant's motor response—and/or accuracy.

It is generally assumed that picture or name characteristics influence at least one specific processing level involved in object naming. The observation of reliable influences of various factors on picture-naming latencies (and/or accuracy) has made several assumptions possible concerning the cognitive processes and the representations acting at specific processing levels. For instance, on the basis of the finding that the age at which object names are acquired (age of acquisition, or AoA; see the Method section) influences naming speed, various hypotheses have been put forward regarding the organization of lexical information (see Juhasz, 2005, for a review)—for instance, the lexical representations corresponding to early-acquired words have lower activation thresholds than do those of late-acquired words, thus making the former easier to retrieve and produce than the latter. Indeed, several normative studies have shown that AoA accounts for a significant part of the variance in picture-naming latencies (e.g., Alario et al., 2004; Barry et al., 1997; Bonin et al., 2002; Liu et al., 2011; Perret & Laganaro, 2013).

From a methodological point of view, normative studies on pictures also make it possible to choose which factors have to be controlled for in factorial experiments. Experimenters face a considerable task when designing factorial experiments, because of the large number of variables that have to be controlled for (Baayen, 2010; Cutler, 1981). For instance, if we again take AoA effects as an example, one may hypothesize that this factor has an influence on the retrieval of orthographic lexemes in handwritten object naming. This hypothesis can be tested by designing a factorial experiment in which electroencephalographic (EEG) activity is recorded (Perret, Bonin, & Laganaro, 2014). Likewise, two sets of pictures can be

contrasted: one set with early-acquired names and another set with late-acquired names. At the same time, other variables that are thought to influence the dependent variable under consideration (e.g., reaction times [RTs], EEG/event-related potentials [ERPs], accuracy) will have to be matched across the two sets of experimental items.

Faced with different variables that can potentially exert an influence on picture-naming performance, it is up to the researcher to decide how to control for them. The most commonly used procedure to control for factors is to match them across experimental conditions. Null-hypothesis significant testing (NHST) is then conducted. If it is found that the two variables do not differ significantly on the variable in question, they are considered to be matched for the nuisance factor (but see Sassenhagen & Alday, 2016, for criticisms of this procedure). However, before matching each nuisance factor across experimental conditions, the researcher has to decide *which* factors need to be controlled for, and normative studies are often used as a guide in this selection process.

Turning to the determinants of object naming, normative studies have revealed that certain factors have a systematic (reliable) influence on naming performance. More particularly, these are name agreement (NA, hereafter) and AoA. NA corresponds to the degree to which participants agree in using a specific label for a drawing (see the Method section) and it is one of the most significant predictors of picture-naming RTs (e.g., Alario et al., 2004; Barry et al., 1997; Bonin et al., 2002; Bonin et al., 2003; Cuetos et al., 1999; Dell'Acqua et al., 2000; Ellis & Morrison, 1998; Johnston et al., 2010; Liu et al., 2011; Nishimoto et al., 2005; Pind & Tryggvadottir, 2002; Severens et al., 2005; Snodgrass & Yuditsky, 1996; Valente et al., 2014; Weekes et al., 2007). When designing a factorial experiment to test the locus of AoA effects, one then has to control for name agreement across experimental conditions (Perret et al., 2014). However, as far as other factors are concerned, there are discrepancies between studies regarding the reliable influences in object naming, such as conceptual familiarity for instance. Conceptual familiarity (see the Method section) is a measure of the degree of physical or mental contact with an object (e.g., *an ashtray* is very familiar for people who smoke) but evidence about the influence of conceptual familiarity in object naming is mixed. In effect, although certain studies have found shorter RTs for highly familiar objects (e.g., Barry et al., 1997; Cuetos et al., 1999; Johnston et al., 2010; Liu et al., 2011; Pind & Tryggvadottir, 2002; Snodgrass & Yuditsky, 1996; Weekes et al., 2007), other studies have failed to find a reliable effect of this variable (e.g., Alario et al., 2004; Bonin et al., 2002; Bonin et al., 2003; Dell'Acqua et al., 2000; Ellis & Morrison, 1998; Nishimoto et al., 2005; Perret & Laganaro, 2013; Severens et al., 2005; Valente et al., 2014). Finally, there are certain variables for which a reliable impact in naming has rarely been reported, as is the case of visual complexity (e.g., Alario et al., 2004, and Ellis &

Morrison, 1998, for studies showing a reliable influence in spoken naming). Visual complexity (see the Method section) is generally measured by asking participants to rate the complexity of the pictures in terms of the number of lines and their intricacy.

Mixed (or null) findings about the reliable impact of certain variables raise difficulties for studies on picture naming. Theoretically, what can be inferred about the cognitive processes and the representations when the variables that are assumed to index their influence are rarely or only inconsistently found to be reliable? Methodologically, researchers have to choose which variables need to be controlled for in a factorial design and doing so is an extremely difficult task. The choice of the variables that have to be controlled for in a factorial design seems to be highly error prone. The first aspect of note is that no confident conclusions can be drawn from a null effect of a variable—that is, a p value greater than .05—on picture-naming performance (Fisher, 1935). As was elegantly described by Dienes (2016; see also Morey, Romeijn, & Rouder, 2016; Rouder, Speckman, Sun, Morey, & Iverson, 2009), the frequentist system involves setting up a model for H_0 alone and trying to reject the null hypothesis. Unfortunately, the inverse approach does not work. A small p value indicates evidence against H_0. However, a large p value does not distinguish between "evidence for H_0" and "not much evidence for anything." The key problem created by this asymmetry is that significance testing cannot provide evidence for the null hypothesis. For instance, no conclusion can be reached from the observation that visual complexity is not reliable in naming speed, because it cannot be excluded that, for instance, the lack of an effect of this variable is a Type II error. The second aspect is practical. After all, why should researchers care about controlling for a factor that is found to be reliable in only half (or even less so, for certain factors) of the published studies? Of course, it is possible to argue that it is always useful to control for some factors even though their impact is weak, in order to avoid some possible confounds. However, it is important to stress that the construction of materials matched on many (and potentially some unnecessary) factors can become an assault course (Cutler, 1981) if the researcher wants to have a reasonable number of items in each experimental condition, in order to prevent a drop in statistical power (Button et al., 2013).

The aim of the present study was to identify which factors should be to be controlled for when building materials for factorial experiments using the picture-naming task, and those that it is (after all) not so important to take into account. To achieve this aim, a meta-analysis on the predictors of naming latencies was conducted. Given that many studies have investigated the determinants of naming speed, we decided to retain for the

analyses only the studies that were the most similar on a set of criteria (e.g., studies using black-and-white pictures, etc.; see the Method section for details). A Bayesian method was used mainly because we had two issues in mind. The first was to identify the factors that account for a significant part of the variance in naming times. In other words, the goal was to identify the factors that researchers cannot ignore when designing a picture-naming factorial study, and that must therefore be controlled for if they are not the subject of the investigation. The second issue is that classical frequentist methods do not allow researchers to make conclusions about the absence of influence of a given variable (Dienes, 2016; Morey et al., 2016). For example, most studies have not reported a reliable influence of visual complexity in object-naming times. Does this variable therefore really have to be controlled for in future studies? The computation of Bayes factors (BF_{10}, hereafter) is an appropriate method for assessing the impact of data on the evaluation of hypotheses. BF_{10} is a ratio between two conditional probabilities (see the Method section). If the ratio is greater than 3, it means that evidence favors the probability to observe H_1; if BF_{10} is less than 1/3, it means that evidence favors the probability to observe H_0; not much evidence is provided either way if BF_{10} is between 1/3 and 3 (Dienes, 2016; Morey et al., 2016).[1] Moreover, BF_{10} indicates whether evidence is weak or strong on a continuum (see Jeffreys, 1961, for a classification). It is important to bear in mind that the Bayesian approach is based on the probability of event distributions and on changes in these distributions due to the outcomes of events (see the Method section). Thinking of Bayesian factors as a dichotomous index of reliability that can be replaced by the p value may be misleading (and the core idea of the Bayesian approach would be lost).

Method

Variables included in the analyses

Even though there are some variations between studies, the influences of the following eight variables in picture-naming latencies have been examined: visual complexity, image agreement, image variability or imageability, conceptual familiarity, name agreement, lexical frequency, AoA, and word length. We now describe how these variables have generally been collected and at what levels of

---

[1] There is a large body of literature on Bayesian analyses. Readers who want to know more about these analyses can consult Kruschke's (2011) book and a recent special issue of *Psychonomic Bulletin & Review* (Vandekerckhove, Rouder, & Kruschke, 2018).
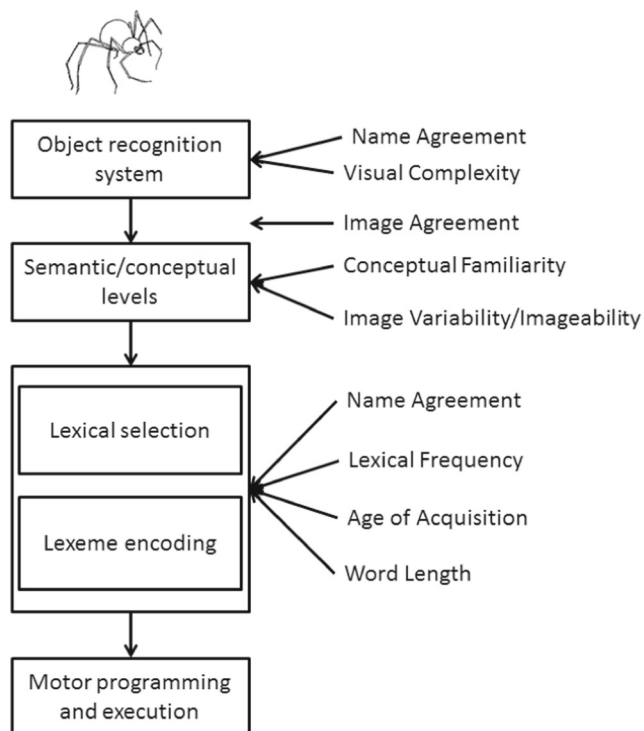
Fig. 1 A model of picture naming, with suggested loci for the different variables investigated in this study

representation involved in object naming they are assumed to occur. We will refer to Fig. 1, which describes a general picture-naming model and indicates the psycholinguistic factors exerting effects on each specific encoding level. Other variables could have been included in the analyses, such as emotional or affective variables (see Hinojosa, Méndez-Bértolo, Carretié, & Pozo, 2010, for an example of a picture-naming study including emotional variables in the analyses of naming times). However, very few picture-naming studies have taken emotional variables into account, and this is obviously a limitation when performing meta-analyses. Moreover, we decided to focus on psycholinguistic factors that have recurrently been investigated in picture naming.

Visual complexity (VC) is rated using Likert scales (generally 1–5 or 1–7 scales) (e.g., 1 = *drawing very simple*, 5 = *drawing very complex*). In this task, participants have to evaluate the complexity of the depicted object in terms of the number of lines and their intricacy. There are objective ways of measuring the visual complexity of pictured objects (e.g., Palumbo, Makin, & Bertamini, 2014; Székely & Bates, 2000), but the great majority of picture-naming studies have used ratings of visual complexity obtained from adults. Visual complexity is generally assumed to index the very first level involved in object naming, namely the perceptual analysis of either the drawing or the photograph corresponding to the object (Fig. 1). At this level, the physical characteristics (shape, surface details, etc.) of

the input are encoded. As far as the influence of VC is concerned, one would expect to find that the more visually complex a picture is, the harder it is to process. However, contrary to this prediction, very few studies (e.g., Alario et al., 2004; Ellis & Morrison, 1998) have found this factor to be a significant predictor of picture-naming speeds.

Image agreement (IA) refers to the degree of matching between a mental image generated from a name (presented visually) and a visual representation of the object referred to by the name. IA is also measured using Likert scales (generally 1–5 or 1–7 scales) with, for instance, a rating of 1 indicating a *very small (or null) degree of matching* and 5 indicating a *very good match*. It is generally assumed that IA captures the similarity between the visual aspect of the objects depicted in a drawing or a photograph and the corresponding stored structural representation (Barry et al., 1997; see Fig. 1). In picture-naming latencies, the influence of IA is predicted to be negative—that is to say, that the RT should decrease as IA increases. Whenever this factor has been included in multiple-regression analyses on naming times, it has generally been found to explain a significant part of the variance.

Image variability (Ivar, or *imageability*) indicates whether a picture label evokes few or many different mental images. Ivar is also measured using Likert scales (1 = *few mental images*, 5 = *many mental images*). *Imageability*, which is closely related to image variability, is also often included as a factor in multiple-regression analyses in picture-naming studies. It is measured using Likert scales by asking participants to rate the ease or difficulty of forming a mental image from the visual presentation of object names. The influence of imageability (or image variability) is assumed to take place at the semantic/conceptual level (Fig. 1). Objects with higher Ivar/imageability scores are easier to produce than objects with lower scores, because the former have "richer" semantic representations (Plaut & Shallice, 1993). Ivar and/or imageability have frequently been reported to be significant predictors of picture-naming RTs (e.g., Alario et al., 2004; Barry et al., 1997; Bonin et al., 2002; Ellis & Morrison, 1998; Perret & Laganaro, 2013).

Conceptual familiarity (Fam) refers to the familiarity of depicted concepts and not to their names. Again it is measured using Likert scales (e.g., 1 indicates a *very unfamiliar concept* and 5 indicates a *very familiar concept*). It is generally assumed that conceptual familiarity influences the ease with which semantic/conceptual representations are contacted (Fig. 1), with shorter RTs being associated with highly familiar items. It has been found that conceptual familiarity influences the picture-naming performance of aphasic patients (e.g., Hirsh & Funnell, 1995). However, no systematic impact of Fam has been observed in the picture-naming performance of healthy participants (e.g., Alario et al., 2004; Bonin et al., 2002; Bonin et al., 2003; Valente et al., 2014, see, however, Snodgrass & Yuditsky, 1996).

Name agreement (NA) refers to the degree of agreement on the use of a specific label for a visual representation corresponding to an object (e.g., a drawing). Two measures of NA are computed: The percentage of participants producing a specific name, and an entropy measure (= h statistic). The latter measure is sometimes preferred, since it takes into account the number of alternative names that are produced for a specific picture. This factor has been shown to be one of the most important predictors of naming speed (e.g., Alario et al., 2004; Bonin et al., 2002; Bonin et al., 2003; Vitkovitch & Tyrrell, 1995), with shorter RTs for items having a high level of name agreement. Two possible loci in object naming have been suggested for NA (Fig. 1). Levelt (2002) assumed that NA has its effects at the level of structural representations. However, another locus for this variable that has been put forward is that of name retrieval (e.g., Barry et al., 1997; Shao, Roelofs, Acheson, & Meyer, 2014; Valente et al., 2014; Vitkovitch & Tyrrell, 1995).

The number of occurrences of a specific lexeme corresponds to *lexical frequency*. This variable is an objective measure and is obtained by counting the number of times a word appears in a specific (spoken and/or written) corpus (e.g., LEXIQUE for French [New, Pallier, Brysbaert, & Ferrand, 2004]; Kučera & Francis, 1967, for English; CELEX for English, Dutch, and German [Baayen, Piepenbrock, & Gulikers, 1993]; etc.). Word frequency is assumed to influence the level at which object name representations are accessed (Fig. 1). Objective word frequency has very often been found to reliably predict object-naming speed (Alario et al., 2004; Barry et al., 1997; Jescheniak & Levelt, 1994; Oldfield & Wingfield, 1965), with shorter naming times being associated with high-frequency picture names.

There has been some debate as to whether word frequency is still a reliable predictor of naming times when AoA is taken into account (e.g., Bonin et al., 2003; Morrison, Chappell, & Ellis, 1997; Morrison, Ellis, & Quinlan, 1992). Indeed, the age at which a picture name is learned corresponds to the *age of acquisition*. Subjective ratings of AoA are obtained by asking adult participants to estimate the age at which they think they learned a word in its oral form (subjective measure). Another way of obtaining AoA measures for words is to analyze the spoken production of children of various ages (objective measures of AoA). RTs are shorter for early-acquired than for late-acquired words. There has been intense debate about the influence of rated AoA on lexical-processing tasks, such as word reading, lexical decision, spelling to dictation, semantic categorization, or spoken/written naming (e.g., Bonin, Barry, Méot, & Chalard, 2004; Bonin, Méot, Mermillod, Ferrand, & Barry, 2009; Pérez, 2007; Zevin & Seidenberg, 2002). Furthermore, the question of exactly how the AoA of words should be measured is still unresolved (see Bonin, Lété, Méot, Roux, & Ferrand, 2016b,

and Lété & Bonin, 2013, for discussions), and we do not intend to take a stance on this issue.[2]

The most important aspect here is that age-limited learning effects are predicted in picture naming, which is the focus of our meta-analysis. Whatever measure ultimately turns out to be the best way to index the age at which words are learned, most researchers in the literature on picture naming have used adult ratings of AoA to index age-limited learning effects, and therefore we also took this variable into account. Another issue related to AoA has been how these effects take place in different lexical tasks, and several hypotheses have addressed the locus of AoA effects: Are AoA effects semantic and/or lexical in nature? At which processing levels do these effects take place in word reading, lexical decision, or picture naming? Also, many hypotheses have been put forward about the precise mechanisms underpinning AoA effects. All these issues remain unresolved (see Johnston & Barry, 2006, and Juhasz, 2005, for reviews). However, as far as picture naming is concerned, even though multiple loci have been proposed for the influence of AoA in picture naming (Johnston & Barry, 2006; Juhasz, 2005), most recent studies have argued that this factor most probably influences the level of lexeme (name) representation (Laganaro & Perret, 2011; Perret et al., 2014; Valente et al., 2014; see Fig. 1). It is important to note that this factor has been found to have a reliable influence in virtually all the picture-naming studies that have examined its influence.

Finally, a measure of *length* is generally included in multiple-regression picture-naming studies. The number of phonemes/syllables, for spoken production, or the number of letters, for handwritten production, is generally used. This factor is thought to index the ease of lexeme encoding (Fig. 1). Although certain studies have reported an influence of

---

[2] Some authors have claimed that AoA ratings provided by adults are valid, whereas others have suggested that the use of AoA scores derived from children's performances (see Johnston & Barry, 2006, for an overview) should be preferred. Other researchers have claimed that both types of AoA measures are problematic, because they are performance variables influenced by a number of other variables, and these researchers have suggested using frequency trajectory instead (Bonin et al., 2004; Zevin & Seidenberg, 2002). The *frequency trajectory* of words corresponds to the variations in exposure to words in print during one's lifetime, from childhood to adulthood (a word's frequency trajectory can be estimated from child and adult frequencies). Frequency trajectory is therefore built from objective word frequency values, and it has been thought to be a better index to use when investigating age-limited learning effects (Bonin et al., 2004; Zevin & Seidenberg, 2002). In particular, certain authors have claimed that when frequency trajectory and better word frequency measures are used, in the form of word frequency values indexing how often words have been encountered throughout a lifetime of reading—that is, the cumulative word frequency—no reliable age-limited learning effects are found in lexical tasks involving quasi-systematic relationships, such as word reading (Bonin et al., 2004; Zevin & Seidenberg, 2002, 2004). However, age-limited learning effects are found in lexical tasks involving arbitrary mappings, such as spoken or written naming (e.g., arbitrary mappings between semantics and names; see Bonin et al., 2004, for empirical evidence using frequency trajectory measures; and Mermillod, Bonin, Méot, Ferrand, & Paindavoine, 2012, for computational evidence).

word length on latencies (e.g., Cuetos et al., 1999; Klapp, Anderson, & Berrian, 1973; Roelofs, 2002; Santiago, MacKay, Palma, & Rho, 2000), this factor has not reached significance in the great majority of other studies (e.g., Bachoud-Lévi, Dupoux, Cohen, & Mehler, 1998; Barry et al., 1997; Dell'Acqua et al., 2000; Snodgrass & Yuditski, 1996).

Studies included in analyses

The studies included in the analyses were selected using the following criteria.

First, the task had to be picture naming and RTs had to be recorded and analyzed. We tried to identify studies in which the naming tasks were very similar so that they involved very similar cognitive processes. The aim was to avoid differences due to variations across modes of cognitive processing.

Second, simultaneous multiple linear regression analyses had to be used for the statistical analyses. Indeed, this was the case in all studies except the Valente et al. (2014) study. In addition, the analyses were all, with the exception of the Valente et al. study, run on aggregate data across items (F_2; Clark, 1973).

Third, the material had to be pictures in black lines on white backgrounds such as those used by Snodgrass and Vanderwart (1980). Recently, Bonin, Méot, Laroche, Bugaiska, and Perret (2017) suggested that the cognitive processes underpinning object naming could, at least in part, intervene differently depending on whether colored or black-and-white pictures are used. The differences in picture formats can add noise in analyses. We therefore decided to exclude studies using grayscale (e.g., Rossion & Pourtois, 2004) or colored pictures (e.g., Bakhtiar, Nilipour, &Weekes, 2013; Tsaparina, Bonin, & Méot, 2011).

Fourth, both written and spoken picture-naming studies were included. In an EEG/ERP study, Perret and Laganaro (2012) observed that the two modes of production started to diverge around 260 ms after picture onset. Following the proposals made by Indefrey (2011), they argued that cognitive processes are common up to the point of lexeme encoding. It is possible to assume that there are differences between the two production modes concerning the effects of AoA and length, respectively (Fig. 1). Perret et al. (2014) have found that AoA has a similar influence on both written and spoken picture naming over the same time scale (from about 260 to 450 ms). For length effects, we decided to include only the studies on oral production in which word length was operationalized in terms of number of phonemes.

Fifth, there is no reason to assume major differences between languages at the level of the cognitive processes involved in object naming. Even though this remains an important theoretical issue, the studies that have explored the determinants of naming speed in different languages have generally found that most of them were shared. Moreover, normative studies generally make use of the same instructions to collect different psycholinguistic norms on pictures or picture names. There can be some differences between languages at the level of the sublexical units that are used in naming, and recent studies have indeed reported certain differences between Chinese and English regarding the role of syllables in spoken word production (e.g., O'Seaghdha, Chen, & Chen, 2010). Nevertheless, in the present study, we did not explore the influence of peripheral/motor factors on RTs. Thus, we included in the analyses studies that were performed with both alphabetic (e.g., English, French, Spanish, etc.) and nonalphabetic (e.g., Chinese, Japanese, etc.) languages.

Likewise, we preselected a set of 18 normative studies. These are as follow: Alario et al. (2004); Barry, Morrison, and Ellis (1997); Bates et al. (2003); Bonin, Chalard, Méot, and Fayol's (2002) spoken and handwritten picture-naming studies, considered separately; Bonin, Peereman, Malardier, Méot, and Chalard (2003); Cuetos, Ellis, and Alvarez (1999); Dell'Acqua, Lotto, and Job (2000); Ellis and Morrison (1998); Johnston, Dent, Humphreys, and Barry (2010); Liu, Hao, Li, and Shu (2011); Nishimoto, Miyawaki, Ueda, Une, and Takahashi (2005); Perret and Laganaro (2013); Pind and Tryggvadottir (2002); Severens, Van Lommel, Ratinckx, and Hartsuiker (2005); Snodgrass and Yuditsky (1996); Valente, Bürki, and Laganaro (2014); and Weekes, Shu, Hao, Liu, and Tan (2007).

However, four studies were excluded from this set. We did not retain Bates et al.'s (2003) study, which explored the predictors of picture naming in seven different languages, because the authors explored only the influences of lexical frequency and length in their analyses. Also, although they included visual complexity, this was estimated on the basis of the size of the JPEG format (Székely & Bates, 2000) and not using a subjective measure (see above). Severens et al.'s (2005) study was not included for similar reasons. That study had only two factors in common—AoA and lexical frequency—with the preselected set. The study by Snodgrass and Yuditsky (1996) was also excluded, because they did not report the *t* test values that are needed to perform Bayesian meta-analyses. Finally, Alario et al.'s (2004) study was not included, because the participants in that study had to name the entire set of pictures twice, and the analyses were reported only for the second naming task. In their EEG/ERP study, Llorens et al. (2014) reported that familiarization and repetition modulate the involvement of cognitive processes in picture naming. Thus, this modulation certainly influences RTs and their predictors to some extent. As a result, Alario et al.'s study is not comparable with other studies in which participants had to name the full set of pictures only once.

Table 1 provides a summary of the 14 studies that were finally selected for the analyses. More information is included in Supplementary Material A. We report the factors and their

Table 1 Summary of the fourteen selected studies and the significance of the eight factors

Studies	Year	Language	N	Type	VC	NA	IA	Ivar	Fam	AoA	Freq	Length
Barry et al.	1997	English	195	Oral	n.s.	✓	✓	✓	✓	✓	✓	n.s.
Bonin et al.	2002	French	203	Oral	n.s.	✓	✓	✓	n.s.	✓	n.s.	n.s.
Bonin et al.	2002	French	201	Written	n.s.	✓	✓	✓	n.s.	✓	n.s.	Letter
Bonin et al.	2003	French	299	Oral	n.s.	H	✓	n.s.	n.s.	✓	✓	n.s.
Cuetos et al.	1999	Spanish	140	Oral	n.s.	✓	✓	–	✓	✓	✓	Syllable
Dell'Acqua et al.	2000	Italian	266	Oral	–	H	–	–	n.s.	✓	–	n.s.
Ellis & Morrison	1998	English British	220	Oral	✓	✓	–	✓	n.s.	✓	✓	n.s.
Johnston et al.	2010	English British	393	Oral	n.s.	✓	–	–	✓	✓	✓	n.s.
Liu et al.	2011	Mandarin Chinese	435	Oral	n.s.	✓	✓	n.s.	✓	✓	n.s.	n.s.
Nishimoto et al.	2005	Japanese	359	Oral	–	✓	–	–	n.s.	✓	n.s.	Mora
Perret & Laganaro	2013	French	120	Written	n.s.	H (n.s.)	✓	✓	n.s.	✓	n.s.	Letter
Pind & Tryggvadottir	2002	Icelandic	171	Oral	–	✓	n.s.	–	✓	✓	n.s.	Letter
Valente et al.	2014	French	100	Oral	n.s.	✓	✓	–	n.s.	✓	n.s.	Syllable
Weekes et al.	2007	Putonghua Chinese	144	Oral	–	✓	✓	–	✓	✓	n.s.	Syllable
% of significance					10%	93%	90%	71%	43%	100%	42%	0%

The percentage significance was computed as the ratio of the number of studies in which the factor is significant divided by the number of studies in which the factor was included. N is the number of pictures used in the study. H corresponds to h-statistic entropy.

significance whenever these were taken into account. It is important to note that the Bayesian meta-analyses were run for NA using the percentage of correct denominations, and not with the h statistic (noted as H in Table 1).

Bayesian meta-analysis

The procedure used derives from that described in the meta-analysis of Rouder and Morey (2011).[3] As we stated above, we chose to use a Bayesian rather than a frequentist approach because the latter does not allow us to test for the presence of invariants (Dienes, 2016; Morey et al., 2016; Rouder et al., 2009). Our aim was to distinguish between the variables that need to be controlled for and those that do not. We must therefore be able to conclude as much about the presence as about the absence of an effect.

For each hypothesis tested (see below), we calculated a Bayes factor as the relative probability of observing the t tests under two competing hypotheses:

$$\mathrm{BF10} = \frac{P(t|\mathrm{H_1})}{P(t|\mathrm{H_0})},$$

with t corresponding to the t tests reported in each study for a given variable, and $\mathrm{H_0}$ and $\mathrm{H_1}$ being the two alternative hypotheses. In Bayesian analyses, the distribution of these conditional probabilities has been referred to as *the posterior*.

---

[3] It should be noted that there is one difference between our study and that of Rouder and Morey (2011): The factors are continuous in our study, whereas they were categorical in Rouder and Morey's study.

Bayesian analyses start with an estimation of the distribution of the probabilities of observing a specific event—for example, the a priori probability of observing a lexical frequency effect. This distribution of probabilities is referred to as *the prior*. The posterior probability is therefore the modification of the prior probability given the observed data—for example, the effect of lexical frequency observed in the literature. The difficulty with the Bayesian approach lies in the estimation of priors. We have followed Rouder and Morey's (2011) proposals based on the so-called Jeffreys–Zellner–Siow (JZS) priors, to honor the contributions of Jeffreys (1961) and of Zellner and Siow (1980), who extended the prior to the class of linear models (Bayarri & Garcia-Donato, 2007). As was explained by Rouder and Morey (p. 686): "Under the JZS priors, we may think of t-statistics as a single piece of datum and the parameter of interest as the effect size δ." In a meta-analytic approach, there is one key property: It is assumed that the true effect size is a constant phenomenon across each experiment. As explained earlier, we have selected studies in which participants had to perform one specific task—that is, picture naming. Our hypothesis is that the cognitive processes underpinning this task are similar across studies as is their impact on RTs. This should translate into effect sizes that are constant across the whole set of studies. The effect size δ was assumed to be zero under the null hypothesis and equal to the Cauchy(r) distribution under the alternative (N. L. Johnson, Kotz, & Balakrishnan, 1994; see Rouder & Morey's, 2011, Appendix Table 3 for details of the mathematical computation). We then computed a Bayes factor for the selected interval against the null. The r-scale parameter controls the scale of the prior distribution (Rouder & Morey, 2011). It represents

one half of the interquartile range of the Cauchy distribution. The Cauchy distribution is a standard distribution for $r = 1$ (wide effect size δ). Each hypothesis was tested using the latter value for the r scale and $\sqrt{2}/2$ (medium effect size δ). We used the conventional approximation guidelines for the strength of evidence provided by Kass and Raftery (1995). The R software (R Core Team, 2015) and the BayesFactor package [meta.ttestBF() function; Morey, Rouder, & Jamil, 2015] were used to compute all Bayes factors.

For each variable, we tested a set of four hypotheses (Table 2). First, we investigated whether the data reported in the 14 studies supported a significant influence of each factor. We therefore computed a Bayes factor from the hypotheses $H_0 \, \delta = 0$ versus $H_1 \, \delta \neq 0$. In a second step, we calculated two new Bayes factors. The first was obtained for $H_0 \, \delta = 0$ versus $H_1 \, \delta \in \,] - \infty; 0 \, [$ —that is, the value of the effect size δ is between minus infinity and 0 (less than 0)—and a second for $H_0 \, \delta = 0$ versus $H_1 \, \delta \in \,] 0; + \infty \, [$ —that is, the value of the effect size δ is between 0 and infinity (greater than 0). We then tested the two directional hypotheses for the alternative hypothesis (H_1). A last Bayes factor was calculated by contrasting the two hypotheses $H_0 \, \delta \in \,] - \infty; 0 \, [$ and $H_1 \, \delta \in \,] 0; + \infty \, [.$[4] With this last Bayes factor, it was possible to test whether the data supported either a negative or a positive influence of the variable. For example, we expected the meta-analysis to support the presence of an AoA effect. However, this is what is predicted by the literature if the BF_{10} is high (>3) with $H_1 \, \delta = \,] - \infty; 0 \, [$, low (<.33) for H1 $\delta = \,] 0; + \infty \, [$ and the Bayes factor obtained from the ratio of the two preceding values is high (>3). Thus, these four Bayes factors indicated whether the meta-analysis argued in favor of the presence or absence of an effect and whether the direction (positive or negative) of the effect was consistent with the predictions derived from the theoretical propositions (effect direction; see Table 2). The data are provided in the Appendix Table 3, and an example of the R codes for AoA can be found in the Supplementary Material B.

Results

Table 2 summarizes all computed Bayes factors. The results were consistent whatever the r-scale values (1 or $\sqrt{2}/2$) were. The Bayes factor did provide very strong ("decisive support," according to Jeffreys, 1961) support for the probability of observing $H_1 \, \delta \neq 0$, given the data reported in the literature (i.e., the posterior probability) for five of the experimental variables. The results of the meta-

____

[4] The latter range was obtained by dividing the BF_{10} obtained for the hypotheses $H_0 \, \delta = 0$ versus $H_1 \, \delta \in \,] - \infty; 0 \, [$ by that obtained for the hypotheses $H_0 \, \delta = 0$ versus $H_1 \, \delta \in \,] 0; + \infty \, [$. Indeed, if we divide two fractions that have the same denominator [e.g., (A/C)/(B/C)], this corresponds to a division of the numerators (A/B).

Table 2 Summary of the Bayes factors computed for the eight factors with the two r scales

Variables	Hypotheses	$r = 1$	$r = \sqrt{2}/2$
Visual complexity	$H_1 \, \delta \neq 0$	0.0184	0.0260
	$H_1 \, \delta \in \,] - \infty; 0 \, [$	0.0060	0.0085
	$H_1 \, \delta \in \,] 0; + \infty \, [$	0.0307	0.0435
	Effect direction	0.1963	0.1962
Name agreement	$H_1 \, \delta \neq 0$	2.27×10^{17}	3.16×10^{17}
	$H_1 \, \delta \in \,] - \infty; 0 \, [$	4.54×10^{17}	6.31×10^{17}
	$H_1 \, \delta \in \,] 0; + \infty \, [$	0.0009	0.0012
	Effect direction	1.89×10^{-21}	1.92×10^{-21}
Image agreement	$H_1 \, \delta \neq 0$	2.81×10^{12}	3.92×10^{12}
	$H_1 \, \delta \in \,] - \infty; 0 \, [$	5.62×10^{12}	7.84×10^{12}
	$H_1 \, \delta \in \,] 0; + \infty \, [$	0.0003	0.0004
	Effect direction	5.24×10^{-17}	5.24×10^{-17}
Image variability	$H_1 \, \delta \neq 0$	160725.8	222977.2
	$H_1 \, \delta \in \,] - \infty; 0 \, [$	321451.6	445954.4
	$H_1 \, \delta \in \,] 0; + \infty \, [$	0.0027	0.0038
	Effect direction	8.40×10^{-9}	8.60×10^{-9}
Conceptual familiarity	$H_1 \, \delta \neq 0$	1,061.7	1,495.4
	$H_1 \, \delta \in \,] - \infty; 0 \, [$	2,123.5	2,990.8
	$H_1 \, \delta \in \,] 0; + \infty \, [$	0.0005	0.0008
	Effect direction	3840359	3840359
Age of acquisition	$H_1 \, \delta \neq 0$	2.64×10^{28}	3.64×10^{28}
	$H_1 \, \delta \in \,] - \infty; 0 \, [$	3.7×10^{-5}	5.11×10^{-5}
	$H_1 \, \delta \in \,] 0; + \infty \, [$	5.28×10^{28}	7.29×10^{28}
	Effect direction	1.43×10^{33}	1.43×10^{33}
Lexical frequency	$H_1 \, \delta \neq 0$	1.87	2.65
	$H_1 \, \delta \in \,] - \infty; 0 \, [$	3.75	5.29
	$H_1 \, \delta \in \,] 0; + \infty \, [$	0.0024	0.0039
	Effect direction	0.0007	0.0067
Length	$H_1 \, \delta \neq 0$	0.0686	0.0969
	$H_1 \, \delta \in \,] - \infty; 0 \, [$	0.0063	0.0089
	$H_1 \, \delta \in \,] 0; + \infty \, [$	0.1309	0.1848
	Effect direction	20.84	20.8

analysis were consistent with the high number of studies reporting significant influences of IA, NA, Ivar/imageability, and AoA—that is, percentages of presence greater than 71%. Mixed findings about the reliable impact of Fam are typical of the difficulties facing studies on picture naming. Indeed, 43% of the studies reported a significant effect of this experimental variable. The Bayes factors provided very strong support ("decisive support," for Jeffreys, 1961) for the posterior probability of observing H_1 for Fam. The Bayes factors very strongly supported the posterior probability of observing a positive influence (H_1 $\delta \in \,] 0; + \infty \, [$) of AoA and Fam, with shorter RTs being associated with highly familiar and early-acquired items. The posterior probability of observing a negative influence ($H_1 \, \delta \in$

] – ∞; 0 [) was supported by the Bayes factors for IA, Ivar/ imageability, and NA, with shorter RTs for higher values.

The Bayes factors argued in favor of the posterior probability of H_0 for VC and length. These results were consistent with the percentages of studies reporting the presence of these effects—10% and 0% for VC and length, respectively. Finally, the Bayes factor for lexical frequency was "barely worth mentioning" (Jeffreys, 1961). Indeed, B_{10} was around 2 for $H_1 \delta \neq 0$, and around 4 for $H_1 \delta \in] - \infty; 0 [$. It seems that these Bayes factors argue in favor of neither the posterior probability of H_0 nor the posterior probability of H_1.

General discussion

Picture naming is an experimental task that has been widely used to study spoken and written word production (e.g., Bonin et al., 2015; Glaser, 1992; C. J. Johnson et al., 1996; Levelt et al., 1999). Thanks to the collection of psycholinguistic norms on both pictures and their names, researchers have found a number of factors that affect naming speed and/or accuracy. The investigation of the influence of several variables on picture-naming performance has made it possible to test important claims about the cognitive processes underlying word production. Given the factors that have been found to reliably affect word production, the list of factors that need to be taken into account when designing a picture-naming experiment has grown steadily. From a methodological point of view, one issue when designing picture-naming experiments has been to control for factors that are thought to reliably affect object naming, but what about the factors that are inconsistently (or rarely) found to be reliable across studies? Should they also be controlled for? As far as word production is concerned, it seems that most researchers have adopted a conservative approach, that is to say they have controlled for the variables that have been controlled for in different studies, even though they have certainly been perfectly aware that some of the controlled variables have rarely been found to have a reliable influence in word production (e.g., visual complexity, see Table 1). We think that proceeding this way renders the process of selecting items for experiments very time-consuming because the list of potential variables to be controlled for can be a very long one, as is the case for picture-naming experiments. It may even become impossible in the future to design any psycholinguistic experiments at all (Cutler, 1981) if, each time a new variable that is found to be reliable (e.g., sensory experience ratings, Juhasz & Yap, 2013) is added to the list of factors and has to be taken into account, while at the same time, those that it seems unnecessary to control for remain in the list. Perhaps one reason for not ceasing to control for variables that have rarely been found to be reliable in picture-naming studies is that classical frequentist methods do not allow researchers to draw conclusions from the absence of influence of a given

variable (Dienes, 2016; Morey et al., 2016). The goal of the present study was to address these issues using a Bayesian meta-analysis and was therefore methodological.

Before continuing with this discussion, there are several points about the interpretation of the findings that need some clarification. As we claimed earlier, a Bayes factor should not be taken as a dichotomous index of reliability that can be replaced by a p value. In spite of what Jeffreys's (1961) classifications might lead us to think, the Bayesian approach is based on the idea of estimating the change in the distribution of the probability of observing a given event when evidence based on previous reports in the literature is taken into account. The present findings therefore have to be interpreted in the light of this approach. First of all, Bayes factors must be considered in the form of ratios. Thus, a BF_{10} of 3.64×10^{28} for AoA (Table 2, r scale = $\sqrt{2}/2$) indicates that taking account of the data available in the literature leads to a change in the probability distribution in favor of the hypothesis $H_1 \delta \neq 0$. Given the data, there is a 3.64×10^{28} greater chance of observing a probability distribution in favor of H_1 than of observing a probability distribution in favor of H_0. Second, this perspective requires us to limit our discussion to certain picture-naming studies—that is, naming studies that have used black-and-white drawings. In effect, the change in the distribution of the probability of observing a given event is based on the evidence that we have taken into account, and only this evidence. One aspect that constrains our interpretations is the fact that we focused on picture-naming studies that used black-and-white drawings as stimuli. Thus, our interpretations and recommendations are limited to such studies. Finally, the Bayesian approach is based on iterative knowledge. This means that the changes in a probability distribution are computed from data that are available at present in the literature, and new findings in the future may change the picture that is described below.

First of all, we found that for most of the variables that have been taken into account in picture-naming studies, namely image agreement, name agreement, image variability/ imageability and AoA, Bayes Factors provided a support in favor of the probability to observe H_1 (i.e., a strong or very strong influence on naming speed) given the data reported in the literature. As we reviewed in the introduction, different views of word production make different claims about the levels at which these variables act. Nevertheless, they all share the idea that image agreement and imageability have their main locus at prelexical stages of word production (Bonin et al., 2002; but see Valente et al., 2014), whereas name agreement and AoA mostly affect postsemantic stages (Perret et al., 2014). When testing hypotheses about the locus of a factor, say AoA in word production, it is clear that the above variables must be taken into account, as otherwise any claim about the locus of AoA and its influence on other stages involved in spoken word production may be flawed. Let us take a recent study conducted by Urooj et al. (2014) as an example. The

authors used magnetoencephalography (MEG) to explore the nature of the influence of AoA on occipital and left anterior temporal cortex activity during object naming. Their findings suggested that there is an initial analysis of object forms in the visual cortex that is not influenced by AoA. Next, a fast-forward sweep of activation occurs from the occipital and left anterior temporal cortex, which brings about stronger activation of semantic representations for early- than for late-acquired names. These findings have important theoretical implications for the role of AoA in spoken word production. However, unfortunately, a close examination of the items used by Urooj et al. reveals that image agreement was not controlled for in their experiment. This is a variable that is assumed to be relevant at the level of structural representations in word production, and as our analysis has revealed, turns out to be a very important factor to take into account.

Second, we found that conceptual familiarity, which measures of the degree of physical or mental contact with an object, and which has not systematically been found to be reliable across studies (for its positive impact on naming times [43% in Table 1], see Barry et al., 1997; Cuetos et al., 1999; Johnston et al., 2010; Liu et al., 2011; Pind & Tryggvadottir, 2002; Snodgrass & Yuditsky, 1996; Weekes et al., 2007; for null-impact: e.g., Alario et al., 2004; Bonin et al., 2002; Bonin et al., 2003; Dell'Acqua et al., 2000; Ellis & Morrison, 1998; Nishimoto et al., 2005; Perret & Laganaro, 2013; Severens et al., 2005; Valente et al., 2014), is nevertheless an important variable to take into account. The change in the distribution of the probabilities in the light of the data reported in the literature argues in favor of an influence of this variable. This is an important and somewhat unexpected finding, which consequently illustrates the strength of the approach used here.

Third, the Bayes factor for lexical frequency was "barely worth mentioning" (Jeffreys, 1961). This finding is also somewhat surprising, given that, in psycholinguistics, word frequency has been claimed to be one of the strongest predictors of processing efficiency (Brysbaert, Mandera, & Keuleers, 2018). In the past, there has been some debate as to whether AoA, lexical frequency, or both factors have influence on picture-naming latencies (Johnston & Barry, 2006). At present, the consensus that has been reached is that both AoA and word frequency are important determinants of naming speed and accuracy (e.g., Bonin et al., 2004). However, the present finding suggests that the influence of lexical frequency in word production should be closely examined in future studies. Once again, such a finding clearly shows how important the Bayesian approach can be. In particular, as recent word recognition studies have suggested (e.g., Brysbaert et al., 2018), some work is certainly needed to decide which of the different word frequency measures that are available best accounts for the variance in naming speed. It is also possible that the influence of lexical frequency in word production

should be examined in relation to vocabulary knowledge, which can vary greatly among individuals. This issue was recently addressed by Mainz, Shao, Brysbaert, and Meyer (2017) in visual word recognition. Interestingly, they found that high-vocabulary individuals exhibited smaller frequency effects in a lexical decision task than low-vocabulary individuals (see also Brysbaert, Lagrou, & Stevens, 2016). Moreover, the mixed-effect reported by Bonin, Laroche, and Perret (2016a) in spelling to dictation seems to be consistent with this hypothesis. These authors showed that the lexical frequency effect was not constant across participants. This suggests that there are variations in the impact of this variable on spelling processing, depending on the participants and, maybe, on the participants' vocabulary knowledge.

Finally, our analyses revealed that the change in the distribution of the probabilities in the light of the data reported in the literature supported the idea of null effects (H_0) for both visual complexity and length, two variables that indeed have rarely been assumed to play a key role in picture naming. At a theoretical level, it must be acknowledged that it is always difficult to make claims about specific mechanisms when the variables assumed to index their influence have rarely or only inconsistently been found to be reliable. Our findings should not therefore be taken to suggest that certain variables, such as word length or visual complexity, never play a role in conceptually driven naming, and that they should no longer be taken into account when building views of word production. In effect, from a computational point of view, it seems difficult to argue that the visual complexity of objects has absolutely no impact on word production because, in picture naming, there is necessarily a level at which some visual characteristics of the pictures are processed. However, our findings make clear that the way that visual complexity has been measured up to now—that is, subjectively, using Likert scales—is certainly problematic, and this should prompt researchers to think of alternative (more reliable) measures of visual complexity. Indeed, there have been some attempts to measure visual complexity *objectively*—for instance, by using the size of the JPEG picture file (see Székely & Bates, 2000).

In conclusion, we think that Bayesian meta-analyses constitute a very useful tool to help researchers gauge the importance of factors in various experimental tasks such as picture naming. We are sure that such analyses will be used more often in the future in several fields of experimental psychology, because they provide useful information for deciding about the factors that should or should not be controlled for when designing experiments.

Appendix

Table 3 Data (*t*-test values) used to run the different Bayes factor with the package BayesFactor

Studies	Year	N	VC	NA	IA	Ivar	FAM	AoA	Freq	Length
Barry et al.	1997	195	1.02	*− 2.9*	*− 1.92*	*− 1.72*	*− 1.92*	*1.82*	*− 3.1*	1.28
Bonin et al. (spoken)	2002	203	− 0.1	*− 3.01*	*− 5.23*	*− 3.83*	0.49	*4.3*	− 0.55	− 1.14
Bonin et al. (written)	2002	201	0.5	*− 1.99*	*− 5.98*	*− 3.43*	0.92	*4.42*	− 0.08	1.5
Bonin et al.	2003	299	1.44	–	*− 3.31*	*− 1.52*	− 0.83	*2.42*	*− 2.95*	− 0.04
Cuetos et al.	1999	140	− 0.47	*− 2.11*	*− 2.04*	–	*− 2.38*	*2.96*	*− 2.47*	–
Dell'Acqua et al.	2000	266	–	–	–	–	0.03	*2.118*	1.65	1.451
Ellis & Morrison	1998	220	*4*	*− 2.07*	–	*− 1.78*	− 1.33	*6.29*	*− 2*	− 0.13
Johnston et al.	2010	393	− 0.278	*− 4.291*	–	–	*− 2.452*	*7.106*	*− 2.04*	0.827
Liu et al.	2011	435	− 0.862	*− 2.326*	*− 2.528*	− 0.974	*− 10.89*	*3.832*	− 0.405	0.868
Nishimoto et al.	2005	359	–	*− 12.5*	–	–	*− 1.49*	*2.05*	− 0.076	–
Perret & Laganaro	2013	120	− 0.08	–	*− 4.13*	*− 3.1*	1.47	*3.88*	*− 1.47*	0.29
Pind & Tryggvadottir	2002	171	–	*− 2.105*	− 0.898	–	*− 3.255*	*5.392*	− 0.542	–
Valente et al.	2014	2,693	− 0.072	*− 4.87*	*− 3.17*	–	0.32	*3.36*	− 0.7	–
Weekes et al.	2007	144	–	*− 4.6*	*− 1.95*	–	*− 3.6*	*2.3*	–	–

N corresponds to the number of observations included in statistical models. The significant *t* tests are in italics.

References

Alario, F.-X., Ferrand, L. Laganaro, M., New, B., Frauenfelder, U. H., & Segui, J. (2004). Predictors of picture naming speed. *Behavior Research Methods, Instruments, & Computers*, *36*, 140–155. doi: https://doi.org/10.3758/BF03195559

Baayen, R. H. (2010). A real experiment is a factorial experiment? *Mental Lexicon*, *5*, 149–157.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1993). The CELEX lexical database (CD ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Bachoud-Lévi, A.-C., Dupoux, E., Cohen, L., & Mehler, J. (1998). Where is the length effect? A cross-linguistic study of speech production. *Journal of Memory and Language*, *39*, 331–346.

Barry, C., Morrison, C. M., & Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency, and name agreement. *Quarterly Journal of Experimental Psychology*, *50A*, 560–585. doi: https://doi.org/10.1080/783663595

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., . . . Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review*, *10*, 344–380. doi: https://doi.org/10.3758/BF03196494

Bayarri, M. J., & Garcia-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, *94*, 135–152.

Bock, K., & Levelt, W. J. M. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), Handbook of psycholinguistics (pp. 945–984). London, UK: Academic Press.

Bonin, P., Barry, C., Méot, A., & Chalard, M. (2004). The influence of age of acquisition in word reading and other tasks: A never ending story? *Journal of Memory and Language*, *50*, 456–476. doi: https://doi.org/10.1016/j.jml.2004.02.001

Bonin, P., Chalard, M., Méot, A., & Fayol, M. (2002). The determinants of spoken and written picture naming latencies. *British Journal of Psychology*, *93*, 89–114. doi: https://doi.org/10.1348/000712602162463

Bonin, P., Laroche, B., & Perret, C. (2016a). Locus of word frequency effects in spelling to dictation: Still at the orthographic level! *Journal of Experimental Psychology: Learning, Memory and Cognition*, *42*, 1814–1820.

Bonin, P., Lété, B., Méot, A., Roux, S., & Ferrand, L. (2016b). At what age did you learn "dog," "harp" and other words that you know? The issue of what adult Age of Acquisition (AoA) estimates really measure. In E. Townsend (Ed.), Learning and memory: Processes, influences and performance (pp. 21–62). New York, NY: Nova Science.

Bonin, P., Méot, A., Lagarrigue, A., & Roux, S. (2015). Written object naming, spelling to dictation, and immediate copying: different tasks, different pathways? *Quarterly Journal of Experimental Psychology*, *68*, 1268–1294.

Bonin, P., Méot, A., Laroche, B., Bugaiska, A., & Perret, C. (2017). The impact of image characteristics on written naming in adults. *Reading and Writing*. Advance online publication. doi: https://doi.org/10.1007/s11145-017-9727-3

Bonin, P., Méot, A., Mermillod, M., Ferrand, L., & Barry, C. (2009). The effects of age of acquisition and frequency trajectory on object naming: Comments on Pérez (2007). *Quarterly Journal of Experimental Psychology*, *62*, 1132–1140. doi: https://doi.org/10.1080/17470210802618850

Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition and naming latencies. *Behavior Research Methods, Instruments, & Computers*, *35*, 158–167. doi: https://doi.org/10.3758/BF03195507

Boukadi, M., Zouaidi, C., & Wilson, M. A. (2016). Norms for name agreement, familiarity, subjective frequency, and imageability for 348 objects names in Tunisian Arabic. *Behavior Research Methods*, *48*, 585–599. doi: https://doi.org/10.3758/s13428-015-0602-3

Brysbaert, M., Lagrou, E., & Stevens, M. (2016). Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingual Language and Cognition*, *20*, 530–548.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: A review update. *Current Directions in Psychological Science*, *27*, 45–50. doi:https://doi.org/10.1177/0963721417727521

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

Cattell, J. M. (1886). The time it takes to see and name objects. *Mind*, *11*, 63–65.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359. doi:https://doi.org/10.1016/S0022-5371(73)80014-3

Cuetos, F., Ellis, A. W., & Alvarez, B. (1999). Naming times for the Snodgrass and Vanderwart pictures in Spanish. *Behavior Research Methods, Instruments, & Computers*, *31*, 650–658. doi:https://doi.org/10.3758/BF03200741

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we able to run any psycholinguistic experiments at all in 1990? *Cognition*, *10*, 65–70. doi:https://doi.org/10.1016/0010-0277(81)90026-3

Dell'Acqua R., Lotto, L., & Job, R. (2000). Naming times and standardized norms for the Italian PD/DPSS set of 266 pictures: Direct comparisons with American, English, French, and Spanish published databases. *Behavior Research Methods, Instruments, & Computers*, *32*, 588–615. doi:https://doi.org/10.3758/BF03200832

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. doi:https://doi.org/10.1016/j.jmp.2015.10.003

Ellis, A. W., & Morrison, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 515–523. doi:https://doi.org/10.1037/0278-7393.24.2.515

Fisher, R. A. (1935). The design of experiments. London, UK: Oliver & Boyd.

Glaser, W. R. (1992). Picture naming. *Cognition*, *42*, 61–105. doi:https://doi.org/10.1016/0010-0277(92)90040-O

Hinojosa, J. A., Méndez-Bértolo, C., Carretié, L., & Pozo, M. A. (2010). Emotion modulates language production during covert picture naming. *Neuropsychologia*, *48*, 1725–1734. doi:https://doi.org/10.1016/j.neuropsychologia.2010.02.020

Hirsh, K. W., & Funnell, E. (1995). Those old, familiar things: Age of acquisition, familiarity, and lexical access in progressive aphasia. *Journal of Neurolinguistics*, *9*, 23–32. doi:https://doi.org/10.1016/0911-6044(95)00003-8

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology*, *2*, 255: 1–17. doi:https://doi.org/10.3389/fpsyg.2010.00255

Jeffreys, H. (1961). Theory of probability. New York, NY: Oxford University Press.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 824–843. doi:https://doi.org/10.1037/0278-7393.20.4.824

Johnson, C. J., Paivio, A., & Clark, J. M. (1996). Cognitive components of picture naming. *Psychological Bulletin*, *120*, 113–139. doi:https://doi.org/10.1037/0033-2909.120.1.113

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). Continuous univariate distributions (2nd ed.). New York, NY: Wiley.

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

Johnston, R. A., Dent, K., Humphreys, G. W., & Barry, C. (2010). British-English norms and naming times for a set of 539 pictures: The role of age of acquisition. *Behavior Research Methods*, *42*, 461–469. doi:https://doi.org/10.3758/BRM.42.2.461

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

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. doi:https://doi.org/10.1080/01621459.1995.10476572

Klapp, S. T., Anderson, W. G., & Berrian, R. W. (1973). Implicit speech in reading, reconsidered. *Journal of Experimental Psychology*, *100*, 368–374.

Kruschke, J. K. (2011). Doing Bayesian data analysis: A tutorial with R and BUGS. Burlington, MA: Academic Press.

Kučera, H., & Francis, W. (1967). Computational analysis of present-day American English. Providence, RI: Brown University Press.

Laganaro, M., & Perret, C. (2011). Comparing electrophysiological correlates of word production in immediate and delayed naming through the analysis of word age of acquisition effects. *Brain Topography*, *24*, 19–29. doi:https://doi.org/10.1007/s10548-010-0162-x

Lété, B., & Bonin, P. (2013). Does frequency trajectory influence word identification? A cross-task comparison. *Quarterly Journal of Experimental Psychology*, *66*, 973–1000.

Levelt, W. J. M. (2002). Picture naming and word frequency: Comments on Alario, Costa and Caramazza. *Language and Cognitive Processes*, *17*, 663–671.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–38, disc. 38–75. doi:https://doi.org/10.1017/S0140525X99001776

Liu, Y., Hao, M., Li, P., & Shu, H. (2011). Timed picture naming norms for mandarin Chinese. *PLoS ONE*, *6*, e16505. doi:https://doi.org/10.1371/journal.prone.0016505

Llorens, A., Trébuchon, A., Riès, S., Liégeois-Chauvel, C., & Alario, F.-X. (2014). How familiarization and repetition modulate the picture naming network. *Brain and Language*, *133*, 47–58.

Mainz, N., Shao, Z., Brysbaert, M., & Meyer, A. S. (2017). Vocabulary knowledge predicts lexical processing: Evidence from a group of participants with diverse educational backgrounds. *Frontiers in Psychology*, *8*, 1164. doi:https://doi.org/10.3389/fpsyg.2017.01164

Mermillod, M., Bonin, P., Méot, A., Ferrand, L., & Paindavoine, M. (2012). Computational evidence that frequency trajectory theory does not oppose but emerges from Age of Acquisition theory. *Cognitive Science*, *36*, 1499–1531.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18.

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes factors for common designs. Retrieved from http://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *Quarterly Journal of Experimental Psychology*, *50A*, 528–559. doi:https://doi.org/10.1080/027249897392017

Morrison, C. M., Ellis, A. W., & Quinlan, P. T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory & Cognition*, *20*, 705–714. doi:https://doi.org/10.3758/BF03202720

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*, 516–524. doi:https://doi.org/10.3758/BF03195598

Nishimoto, T., Miyawaki, K., Ueda, U., Une, Y., & Takahashi, M. (2005). Japanese normative set of 359 pictures. *Behavior Research Methods*, *37*, 398–416. doi:https://doi.org/10.3758/BF03192709

O'Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition*, *115*, 282–302.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*, 273–281.

Palumbo, L., Makin, A. D. J., & Bertamini, M. (2014). Examining visual complexity and its influence on perceived duration. *Journal of Vision*, *14*(14), 3:1–18. doi:https://doi.org/10.1167/14.14.3

Pérez, M. A. (2007). Age of acquisition persists as the main factor in picture naming when cumulative frequency and frequency trajectory are controlled. *Quarterly Journal of Experimental Psychology*, *60*, 32–42.

Perret, C. Bonin, P., & Laganaro, M. (2014). Exploring the multiple-level hypothesis of AoA effects in spoken and written picture naming using a topographic ERP analysis. *Brain and Language*, *135*, 20–31.

Perret, C., & Laganaro, M. (2012). Comparison of electrophysiological correlates of writing and speaking: A topographic ERP analysis. *Brain Topography*, *25*, 64–72.

Perret, C., & Laganaro, M. (2013). Why are written naming latencies (not) longer than spoken naming? *Reading and Writing*, *26*, 225–239.

Pind, J., & Tryggvadottir, H. B. (2002). Determinants of picture naming times in Icelandic. *Scandinavian Journal of Psychology*, *43*, 221–226. doi:https://doi.org/10.1111/1467-9450.00290

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377–500. doi:https://doi.org/10.1080/02643299308253469

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Roelofs, A. (2002). Syllable structure effects turn out to be word length effects: Comment on Santiago et al. (2000). *Language and Cognitive Processes*, *17*, 1–13. doi:https://doi.org/10.1080/01690960042000139

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*, 217–236. doi:https://doi.org/10.1068/p5117

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682–689. doi:https://doi.org/10.3758/s13423-011-0088-7

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* test for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. doi:https://doi.org/10.3758/PBR.16.2.225

Sanfeliu, M. C., & Fernandez, A. (1996). A set of 254 Snodgrass–Vanderwart pictures standardized for Spanish: Norms for name agreement, image agreement, familiarity, and visual complexity.

Behavior Research Methods, Instruments, & Computers, *28*, 537–555. doi:https://doi.org/10.3758/BF03200541

Santiago, J., MacKay, D. G., Palma, A., & Rho, C. (2000). Sequential activation processes in producing words and syllables: Evidence from picture naming. *Language and Cognitive Processes*, *15*, 1–44.

Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance parameter control with the null-hypothesis rejection tests. *Brain and Language*, *162*, 42–45.

Severens, E., Van Lommel, S., Ratinckx, E., & Hartsuiker, R. J. (2005). Timed picture naming norms for 590 pictures in Dutch. *Acta Psychologica*, *119*, 159–187. doi:https://doi.org/10.1016/j.actpsy.2005.01.002

Shao, Z., Roelofs, A., Acheson, D. J., & Meyer, A. S. (2014). Electrophysiological evidence that inhibition supports lexical selection in picture naming. *Brain Research*, *1586*, 130–142. doi:https://doi.org/10.1016/j.brainres.2014.07.009

Snodgrass, J. G., & Vanderwart, M. (1980). A Standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 174–215. doi:https://doi.org/10.1037/0278-7393.6.2.174

Snodgrass, J. G., & Yuditsky, T. (1996). Naming times for the Snodgrass and Vanderwart pictures. *Behavior Research Methods, Instruments & Computers*, *28*, 516–536. doi:https://doi.org/10.3758/BF03200540

Székely, A., & Bates, E. (2000). Objective visual complexity as a variable in studies of picture naming. *CRL Newsletter*, *12*, 3–33.

Tsaparina, D., Bonin, P., & Méot, A. (2011). Russian norms for name agreement, image agreement for the colorized version of the Snodgrass and Vanderwart pictures and age of acquisition, conceptual familiarity, and imageability scores for the modal object names. *Behavior Research Methods*, *43*, 1085–1099. doi:https://doi.org/10.3758/s13428-011-0121-9

Urooj, U., Cornelissen, P. L., Simpson, M. I. G., Wheat, K. L., Woods, W., Barca, L., & Ellis, A. W. (2014). Interactions between visual and semantic processing during object recognition revealed by modulatory effects of age of acquisition. *NeuroImage*, *87*, 252–264.

Valente, A., Bürki, A., & Laganaro, M. (2014). ERP correlates of word production predictors in picture naming: a trial by trial multiple regression analysis from stimulus onset to response. *Frontiers in Neuroscience*, *8*, 390. doi:https://doi.org/10.3389/fnins.2014.00390

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*, 1–4. doi:https://doi.org/10.3758/s13423-018-1443-8

Vitkovitch, M., & Tyrrell, L. (1995). Sources of disagreement in object naming. *Quarterly Journal of Experimental Psychology*, *48A*, 822–848. doi:https://doi.org/10.1080/14640749508401419

Weekes, B. S., Shu, H., Hao, M., Liu, Y, & Tan, L. H. (2007). Predictors of timed picture naming in Chinese. *Behavior Research Methods*, *39*, 335–342. doi:https://doi.org/10.3758/BF03193165

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), Bayesian statistics: Proceedings of the First International Meeting, held in Valencia (Spain) (pp. 585–603). Valencia, Spain: University of Valencia.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, *47*, 1–29. doi:https://doi.org/10.1006/jmla.2001.2834

Zevin, J. D., & Seidenberg, M. S. (2004). Age-of-acquisition effects in reading aloud: tests of cumulative frequency and frequency trajectory. *Memory & Cognition*, *32*, 31–38. doi:https://doi.org/10.3758/BF03195818