Review of Philosophy and Psychology (2021) 12:745–774
https://doi.org/10.1007/s13164-020-00503-8

Intuitions About the Reference of Proper Names: a Meta-Analysis

Noah van Dongen[1] · **Matteo Colombo**[2] · **Felipe Romero**[3] · **Jan Sprenger**[1]

Published online: 20 September 2020
© The Author(s) 2020

Abstract
The finding that intuitions about the reference of proper names vary cross-culturally (Machery et al. *Cognition* 92: 1–12. 2004) was one of the early milestones in experimental philosophy. Many follow-up studies investigated the scope and magnitude of such cross-cultural effects, but our paper provides the first systematic meta-analysis of studies replicating (Machery et al. *Cognition* 92: 1–12. 2004). In the light of our results, we assess the existence and significance of cross-cultural effects for intuitions about the reference of proper names.

Keywords Semantic intuitions · Theory of reference · Proper names · Cross-cultural psychology · Meta-analysis · Experimental philosophy

1 Introduction

Most people who have heard of the Italian mathematician Giuseppe Peano credit him with inventing the standard axioms of arithmetic. This is all they associate with the name "Peano". Yet, the axioms were invented by the German mathematician Richard Dedekind, and Peano published a simplified version only afterwards. If people identify Peano only by the description "the inventor of the standard axioms of arithmetic", to whom are they referring when they use the name "Peano"? To Peano or to Dedekind? And more generally, what kind of meaning must people associate to a proper name like "Peano" in order to be competent users of that name?

Supplements: All material, including data and analysis code, is available on the OSF project page: https://osf.io/et86f/.

✉ Matteo Colombo
m.colombo@uvt.nl

[1] Department of Philosophy, Turin University, Turin, Italy

[2] Tilburg center for Logic, Ethics and Philosophy of Science, Tilburg University, Tilburg, The Netherlands

[3] Department of Philosophy, Groningen University, Groningen, The Netherlands

In philosophy of language, there are two main classes of theories about the meaning of proper names: descriptivist theories and causal-historical theories. According to descriptivist theories (Frege 1892; Russell 1905; Searle 1958), proper names have definite descriptions as their meaning. The idea is that a proper name can refer to a person only via the *descriptive properties that users of the name associate with it*. Thus, people who identify Peano only by the description "the inventor of the standard axioms of arithmetic" would actually refer to Dedekind when they use the name "Peano". After all, it is Dedekind who uniquely satisfies that description.

According to causal-historical theories (Kripke 1980), proper names do not imply any descriptive property of the individuals to which they refer. Proper names refer directly to their bearers without being essentially associated with any descriptive properties of an individual. People processing a proper name like "Peano" certainly rely on some mental representations of certain descriptive properties, but these representations play no role in determining the meaning of "Peano". Instead, what is crucial for determining the meaning of a proper name is its *causal history*. All uses of the name that causally derive from an original act of baptism refer to the individual originally baptized with that name. Even if people falsely associate the description "the inventor of the standard axioms of arithmetic" with Peano, the proper name "Peano" actually refers to Peano. This is because of a relevant causal chain between the original act of baptism for Giuseppe Peano, and people's usage of the name "Peano".

One influential argument for why proper names cannot be semantically equivalent to definite descriptions is that the referent of a proper name 'is stipulated to be a single object whether we are speaking of the actual world or a counterfactual situation' ((Kripke 1980), p. 21). In contrast, definite descriptions can refer to different individuals in counterfactual situations.

Kripke illustrates this argument with an example analogous to the Peano case above ((Kripke 1980), pp. 83 ff.). Suppose that the only description you associate with the name "Gödel" is "the mathematician who proved the incompleteness of arithmetic". Suppose that you discover that a certain Mr Schmidt rather than Gödel actually proved the incompleteness of arithmetic. If the name "Gödel" is semantically equivalent to the definite description "the mathematician who proved the incompleteness of arithmetic", then you are committed to the conclusion that the name "Gödel" actually refers to Mr Schmidt. But, of course, this conclusion is false, which suggests that descriptivist theories of proper names are false.

In a landmark study, Machery et al. (2004) used this Gödel case, and three other similar vignettes, to explore the question of whether Kripke's and other Anglophone speakers' judgments about the referents of proper names are cross-culturally robust. (Machery et al. 2004) found that Westerners tend to have intuitions in line with Kripke's causal-historical theory, while East Asians' intuitions tend to agree with the descriptivist theory. On the basis of this finding, (Machery et al. 2004) reasoned that if people's judgments about the meaning of a proper name are systematically influenced by demographic variables like their culture, then Kripke's and other Anglophone speakers' semantic judgments cannot constitute reliable evidence for a theory of proper names.

Several follow-up studies have extended and probed the original result, and embedded it into a broader methodological discourse about the use of intuitions about particular cases as evidence for philosophical theories e.g.,(Machery et al. 2009; Lam 2010; Sytsma and Livengood 2011; Machery 2017; Cova et al. 2018). (Machery et al. 2004)'s effect has been found to show significant variation both within and across studies, but no convincing explanation has been offered for when and why we should expect cross-cultural variation in semantic intuitions. It might be that scenarios like the Peano or Gödel cases above are not the right tools for eliciting people's semantic intuitions. Or, it might be that for making semantic judgments people have access to multiple cognitive strategies, which include both causal-historical and descriptivist factors, and shift between them depending on one's ease to take the perspective of the speaker, background knowledge, audience, and purpose of communication (Genone and Lombrozo 2012).

Philosophers are not the only ones interested in proper names. Proper names have also been studied by several other disciplines, including linguistics, psychology, neuroscience, and anthropology.

Linguists agree that proper names are mainly used to identify individuals uniquely; but the referential and connotative use of a proper name depends on the communicative intentions the speaker want to convey to the hearer, the knowledge frame shared by the interlocutors, and their discourse-relative perspective (Dancygier 2009; Dancygier and Vandelanotte 2017; Marmaridou 2000).

Cognitive psychologists focus on the mental representations of the meanings of proper names, and in particular on differences between the cognitive processing of proper names and common nouns (Sophia and Marmaridou 1989; Valentine et al. 1996). Much of this research provides evidence that proper names and common nouns are associated with different mental representations, and that they are processed differently. It has been found that, generally, people from different cultures judge more quickly if a word is a proper name as opposed to a common noun (Müller 2010), and that they retrieve the meaning of a proper name from memory with more difficulty compared to common nouns (Proverbio et al. 2009; Wang et al. 2016). Additional evidence for these processing differences between proper names and common nouns comes from neuropsychological studies, which have shown a double dissociation between retrieval of proper names and retrieval of common nouns (Yen 2006; Semenza 2006).

Finally, for anthropologists, the meaning of proper names depends on principles governing naming practices across cultures (Vom Bruck and Bodenhorn 2006; Bright 2003). Proper names have been found to fulfil two main cultural functions: to identify their bearers differentiating them from other individuals, and to classify them in terms of their parental, economic, ethnic or geographical group. These two functions can trade off, since the more a proper name differentiates its bearer from other individuals, the less social information it carries. Focusing on these trade-offs, (Alford 1988) examined the use of proper names in sixty cultures around the world to better understand the social and communicative situations where one function is more prevalent than the other.

Given this rich and varied literature, it is noteworthy that the philosophical literature studying cross-cultural variation of proper names has overlooked the connection

to studies in other disciplines about how people use, memorize, recall, and mentally represent the meaning of proper names.[1] At the very least, studies from other disciplines, which employ various methods to study proper names, would provide experimental philosophers with independent evidence to probe the robustness of putative cross-cultural effects.

Locating experimental philosophers' work in the wider scientific context highlights the general importance of the questions we set out to address with our meta-analysis in this paper. For example, to what extent are the demographic effects observed by experimental philosophers large and interpretable? Do they depend on superficial features of the experimental material and design experimental philosophers have been using? Or do they indicate variation in the mechanisms and representations, which different communities of speakers would recruit to make sense of proper names? Answering these questions is not only important for philosophers; it will also clarify whether researchers from different disciplines are warranted to rely on the demographic effect Machery et al. (2004) originally found, and on simple vignettes like the Gödel case for studying proper names.

A relatively large and interpretable meta-analytic effect will be convincing reason that the cross-cultural effect is not an artefact of particular vignettes. The finding that a descriptivist theory of proper names predicts East Asians' intuitions, while a causal-historical theory predicts Westerners' intuitions, will motivate new hypotheses cognitive psychologists, and neuropsychologists could test. For example, as noted above, psychologists and neuropsychologists have proposed that the retrieval of proper names is particularly difficult compared to common nouns. One reason for this hypothesis is that proper names would be detached from the semantic network representing descriptive or biographical information (e.g., Semenza, 2006). But, if East Asian speakers have descriptivist intuitions, then their retrieval of proper names should be easier compared to the retrieval of proper names exhibited by Western speakers, whose mental representations of proper names would be detached from descriptive information.

A small, noisy, and hard-to-interpret meta-analytic effect will give us reason to either call into question the idea that demographic factors play a substantial role in semantic intuitions, or doubt that the vignettes philosophers have been using to elicit such intuitions are reliable tools. Either way, this finding should motivate experimental philosophers interested in the semantics of proper names to pay closer attention to relevant anthropological evidence about different functions of proper names, as well as to the experimental designs cognitive psychologists and linguists have been using to study proper names.

So, central to our meta-analysis are the questions of whether Westerners' and East Asians' intuitions about the reference of person names are in line with either the causal-historical or the descriptivist theory (Hypotheses 1a and 2a), and how the two

[1] This is the more surprising because the original study and hypotheses tested by (Machery et al. 2004) were motivated by Richard Nisbett and his collaborators' studies in cultural psychology (Nisbett et al. 2001). These studies suggested that different cultural groups exhibit different styles of thought: while "holistic thought" would be prevalent among East Asians, "analytic thought" would be predominant among Westerners.

populations compare in this respect (Hypothesis 3a).[2] We evaluate these hypotheses on the basis of a set of empirical studies, which rely on probes similar to the Gödel and Peano cases we have described above. To answer these questions, we test three hypotheses:

Hypothesis 1a	The majority of Westerners have causal-historical intuitions.
Hypothesis 2a	The majority of East Asians have descriptivist intuitions.
Hypothesis 3a	Westerners are more likely than East Asians to have causal-historical intuitions.

There are two types of vignettes in the literature we aggregate: first, the *Gödel probe*, which is structurally identical to our introductory Peano/Dedekind example, and the *Jonah probe*, where the properties associated with a proper name gradually shift over time. In the Gödel probe both the causal-historical and the descriptivist theory single out a specific referent of the proper name, while in the Jonah probe the proper name does not refer at all according to the descriptivist account. Since both probes follow the same experimental design (exposure to vignette, binary response, two theories with different predictions, etc.), the results can be aggregated to test hypotheses 1a–3a. Moreover, since the Gödel probe has particular significance in the philosophical literature (? [, going back to)]Kripke1980, and since the original study found a notable effect of cross-cultural variation only for Gödel probes, but not for Jonah probes (Machery et al. 2004), we also test the following three hypotheses:

Hypothesis 1b	The majority of Westerners have causal-historical intuitions in the Gödel probes.
Hypothesis 2b	The majority of East Asians have descriptivist intuitions in Gödel probes.
Hypothesis 3b	Westerners are more likely than East Asians to have causal-historical intuitions in Gödel probes.

One final note before we describe our meta-analysis in detail. Different studies in the literature originated from (Machery et al. 2004)'s work differ in the samples of Westerners and East Asians. (Machery et al. 2004) tested participants in the United States and Hong Kong. Follow-up studies use for example Dutch participants (Cova et al. 2018) or French participants (Machery et al. 2009). We note these differences in detail in Section 3.1. Nonetheless, we formulated the hypotheses in a general form as a comparison between Westerners and East Asians for two reasons. First, we intended to study the effect as reported in (Machery et al. 2004) and hence we followed their design. Second, we did not have theoretical reasons to exclude e.g., Dutch or French from the group of Westerners or e.g., Japanese from the group of East Asians.

[2]Compare the following quote from the original paper (? [)p.B5]MacheryEtAl2004: "W[esterner]s would be more likely to respond in accordance with causal-historical accounts of reference, while E[ast]A[sian]s would be more likely to respond in accordance with descriptivist accounts of reference."

2 Material and Methods

2.1 Data Sources and Searches

We conducted a comprehensive literature search of studies on cross-cultural semantic intuitions. To find replications of the original experiment, we started with the Google Scholar list of all papers citing (Machery et al. 2004) and checked whether they contained experimental data. Our search was aided by a list of known replications that we obtained via e-mail from Édouard Machery (=the first author of the original study). We used this search strategy because any replication of Machery et al. (2004) would, in virtue of its aims and scope, refer to the original paper.

2.2 Study Selection

Studies were eligible if they were published or publicly available in English, and used a design sufficiently similar to the original study. Specifically, eligible studies had to contain results from experiments featuring East Asian and/or Western participants with one or more binary-choice probe. The binary-choice answer to such probes had to correspond to the descriptivist and the causal-historical theory of reference, respectively.

2.3 Data Extraction

The eligible studies were classified by two teams, each of which included two authors of this paper. For each study, team members independently extracted the name of the first author, the year of publication, and the data of the probe responses. Per study and probe, they independently extracted data on the type of the probe (e.g., Gödel, Jonah, etc.), the number of Western and East Asian participants, total sample size, the number of causal-historical responses (per subgroup and in total), and deviations from the original design, such as language of the probe, phrasing of the question, and phrasing of the answers. Disagreements were resolved by consensus.[3]

2.4 Data Synthesis and Analysis

We carried out a confirmatory (Section 3.2) and an exploratory analysis (Section 3.3), as well as a quality appraisal Section 3.4 and Appendix C). A confirmatory analysis assessed the overall meta-analytic evidence and its generalisability for the main hypotheses tested in Machery et al. (2004) (Hypotheses 1a–3a) and the meta-analytic evidence for the particular Gödel probes (Hypotheses 1b-3b) for which Machery et al. (2004) reported positive results. Specifically, for Hypotheses 1b-3b, only Gödel probes were used, which did not deviate from the design of the original study (i.e., direct replications). The analyses were conducted on the level of the individual probes within a study, because not all studies had the same (number of) probes. We used

[3]The full data spreadsheet is available online at https://osf.io/et86f/.

a multilevel random effects (RE) model to capture the hierarchical structure (i.e., probes within studies) of the data and the inter-study dependencies between probes via the 'metafor' package in R (Viechtbauer 2010).

In total, we performed six confirmatory meta-analyses, one per hypothesis. Specifically, we calculated summary proportion estimates of Causal/Historical response for single cultural group responses (Hypotheses 1a/b, 2a/b) with a restricted maximum-likelihood heterogeneity estimator[4] to model the random effects. For the probes that compare Westerners to East Asians (Hypotheses 3a and 3b), we calculated summary Relative Risk ratios (RR), that is, the quotient of the proportions of Causal/Historical responses in both groups: If A and B are the two groups and p_A and p_B are the proportions of event E in the groups (in our case: Causal/Historical responses), then the relative risk ratio is defined as $\mathrm{RR}_E(A, B) = p_A/p_B$.[5] The extent of heterogeneity between probes was assessed by the I^2 measure (indicating the percentage of total variance due to between-probe variance; Higgins et al. (2003) and Ioannidis et al. (2007)). In addition to the 95% confidence intervals (CIs) for the unknown parameters (i.e., the proportion of causal-historical responses and the RR between populations), we calculated the 95% prediction intervals (PI). Unlike CIs, which are based on compatibility of the observed data with an unknown parameter value, PIs predict the distribution of future data points by taking into account inter-study variability. Therefore, they are better suited to express a plausible range of values for the next conducted study, and to predict whether effects are likely to replicate (Higgins et al. 2009; Riley et al. 2011).

The exploratory analysis focused on the observed variance in effect-size between the various probes; that is, it aimed to identify factors that would explain why studies often deliver heterogeneous results. Specifically, we repeated the tests performed for Hypotheses 1a–3a with a meta-regression analysis where deviations from the original design were included as moderator variables. We ignored the hierarchical structure of the data because we were interested in explaining variance by the general design factors.[6]

3 Results

3.1 Studies and Data

Via Google Scholar, we identified 482 records published in 2017 or earlier that cite (Machery et al. 2004). Using the search criteria described in Section 2.1, we ended

[4]This estimator is approximately unbiased and efficient (Viechtbauer 2005; Raudenbush 2009), and is the software package's default (Viechtbauer 2010).

[5]Relative Risk is a well-probed measures of association between categorical data that is easy to interpret and using a different effect-size measure like Odds Ratio does not change the inference (Higgins and Green 2011).

[6]Taking account of hierarchical structures allows one to assess the associations between level, but ignoring this structure has a negligible effect on the main effect estimates (O'Mara 2008).

up with 15 potential studies. Four studies were added in addition to our initial search results, resulting in a total of 19 studies.[7]

Eventually, from this set of 19 studies, 13 were included and 6 were excluded. Reasons for exclusion were the following: [1] missing data, [2] culturally mixed samples, and [3] structure of the data (i.e., data from non-binary questions, compare Section 2.2). See Appendix A for the list of excluded papers and their reason for exclusion. In addition, the Indian sample from Machery 2009 was excluded, because they are considered to be an South-Asian instead of East-Asian. Combined, these studies tested 61 probes on 4691 participants, who produced a total 8959 binary responses. Of these probes, 35 tested both Western and East Asian samples [median sample size: 181]; 15 probes tested Western samples [median sample size: 60]; and 11 probes tested East Asian samples [median sample size: 211]. Table 1 provides an overview of the number of probes per study and their characteristics. Most Western samples were from the United States of America. Exceptions were 'Machery, 2009' (France), 'Cova, 2019' (The Netherlands), and 'Colombo, n.p.' (The Netherlands, USA, England, Germany, and Italy). In general, the East-Asian samples were from Hong Kong. Exceptions were 'Machery, 2009' (Mongolia), 'Sytsman, 2015' (Japan), 'Kazaki, 2017' (Japan), 'Izumi, 2018' (Japan), and 'Colombo, n.p.' (Hong Kong and China). We identified ten factors on which probes could deviate from the original design (see Appendix B), from the language in which probes were presented to the phrasing of the question that participants were asked.[8]

In particular, the studies in our dataset did not always use the same phrasing for eliciting a judgment on causal-historical versus descriptivist intuitions. Most studies followed the original design by (Machery et al. 2004) and asked participants the following question: Who is the person that the vignette character John is "talking about" when using the proper name "Peano"? By contrast, (Sytsma and Livengood 2011) suggest rephrasing the question in the following way: Who would you *take* is the person that John is talking about when using the proper name "Peano"? (this modification is known as the "clarified narrator's perspective", see also Section 3.3), and (Machery et al. 2009) asked for the *truth value* of sentences such as "Peano discovered the incompleteness of mathematics". Similarly, answers to the binary question were sometimes phrased as descriptions (e.g., in the original study) and sometimes as bare noun phrases without further explanations e.g., in (Lam 2010). For the purpose of the meta-analysis, we treated all these dependent variables as answering the same question, independently of the exact phrasing. However, for testing Hypothesis 1b-3b, we focused on direct replications of the original Gödel probe and excluded variations of the design such as Sytsma and Livengood's "clarified narrator's perspective". These different ways of analyzing the data allow us to answer two different questions: first, the question of the replicability and internal validity of the original

[7] In personal communication, Édouard Machery suggested to add the original results reported in (Machery et al. 2015) and a forthcoming by (Izumi et al. 2018). In addition, we included the results of our replication attempt reported in (Cova et al. 2018) and an online replication attempt carried out via Qualtrics, whose details are also available online at https://osf.io/et86f/.

[8] The complete list of the design deviation factors and their values is given in the data extraction plan, which is available online in the additional materials file.

Table 1 Overview of the Studies and Their Characteristics

First author	Year	Study Type	Study Design	Number of probes	Sample Size	Design Deviations
Machery	2004	Comparison	Within-subjects	4 (Gödel, Jonah)	72	No (original study)
Lam	2009	Comparison	Within-subjects	3 (Shakespeare, Julius)	69	Yes (2 deviations)
Machery	2009	Comparison	Between-subjects	2 (Gödel, Shakespeare)	144	Yes (2 deviations)
Machery	2010	Asians only	Single probe	1 (Gödel)	19	Yes (1 deviation)
		Comparison	Single probe	1 (Gödel)	153	Yes (1 deviation)
Sytsma	2011	Westerners only	Within-subjects	3 (Gödel)	35	Yes (1 deviation)
		Westerners only	Between-subjects	3 (Gödel)	223	Yes (1 deviation)
Beebe	2015	Westerners only	Between-subjects	6 (Satyricon)	304	Yes (1 deviation)
		Westerners only	Between-subjects	3 (Satyricon)	180	Yes (2 deviations
		Comparison	Within-subjects	4 (Satyricon)	177	Yes (1 deviation)
Machery	2015	Asians only	Single probe	1 (Gödel)	65	Yes (2 deviations)
		Comparison	Within-subjects	2 (Gödel)	129	Yes (2 deviations)
Sytsma	2015	Comparison	Between-subjects	3 (Gödel)	621	Yes (2 deviations)
Beebe	2016	Comparison	Within-subjects	4 (Gödel, Jonah)	207	Yes (1 deviation)
	2016	Comparison	Within-subjects	4 (Gödel, Jonah)	406	Yes (1 deviation)
Kazaki	2017	Asians only	Within-subjects	2 (Gödel)	231	Yes (3 deviations)
Izumi	2018	Asians only	Between-subjects	7 (Gödel)	1283	Yes (2 deviations)
Cova	2018	Comparison	Within-subjects	4 (Gödel, Jonah)	181	No (only sample deviation)
Colombo	n.p.	Comparison	Within-subjects	4 (Gödel, Jonah)	192	No (only sample deviation)

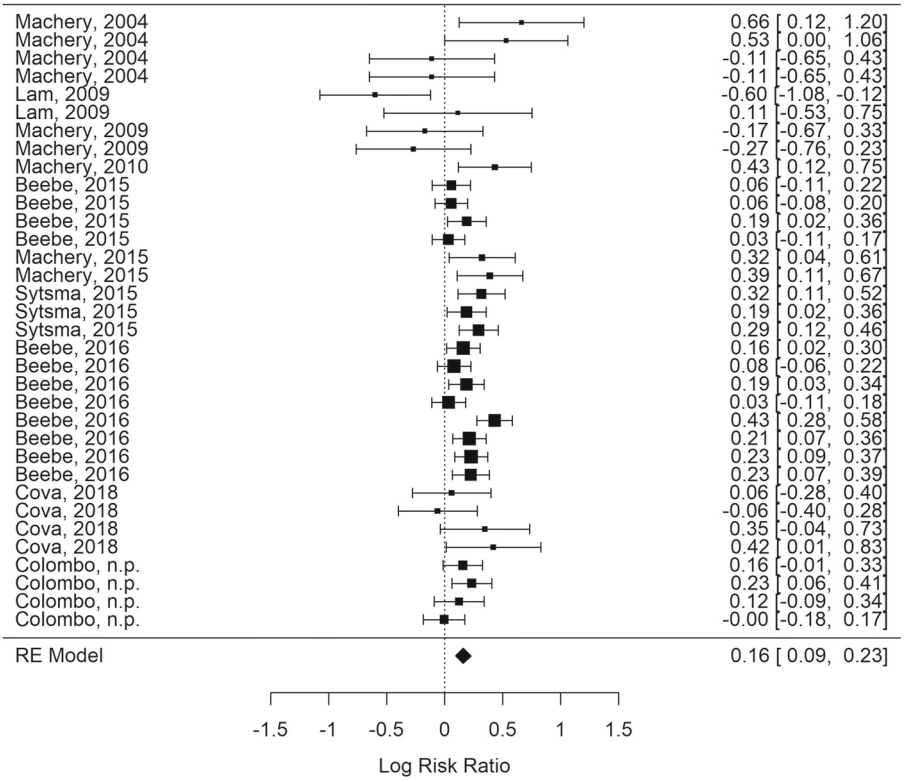

Fig. 1 Forest plot for Hypotheses 3a, showing distribution of effect sizes and confidence intervals

effect observed by Machery et al., and second, the generality and external validity of the observed effect to a wider hypothesis about the reference of proper names.

3.2 Meta-Analytic Estimates

The results per hypothesis are displayed as forest plots in Figs. 1 and 2[9] and summarized below. The analyses provide confirmatory evidence for Hypotheses 1a, 2b, 3a and 3b in the sense that the confidence interval for the unknown parameter consists only of values in the expected direction, and excludes the hypothesis of zero effect—that is, a relative risk ratio of RR = 1 for hypotheses 3a and 3b, and a 50% proportion of Causal/Historical responses for hypotheses 1a through 2b. Forest plots for all six hypotheses are given in Figs. 1 and 2. Below is a precise statement of our results.

[9]These plots show the distribution of effect-sizes (location, confidence intervals, weight in the meta-analysis) of the probes that were used for the analysis, their summary meta-analytic effect-size (diamond at the bottom), and its confidence interval (width of the diamond). The forest plots for Hypothesis 1a and 2 can be found in the online supplementary files. Please note that the plots show log-transformed effect-sizes. The untransformed effect-sizes are reported in the main text.

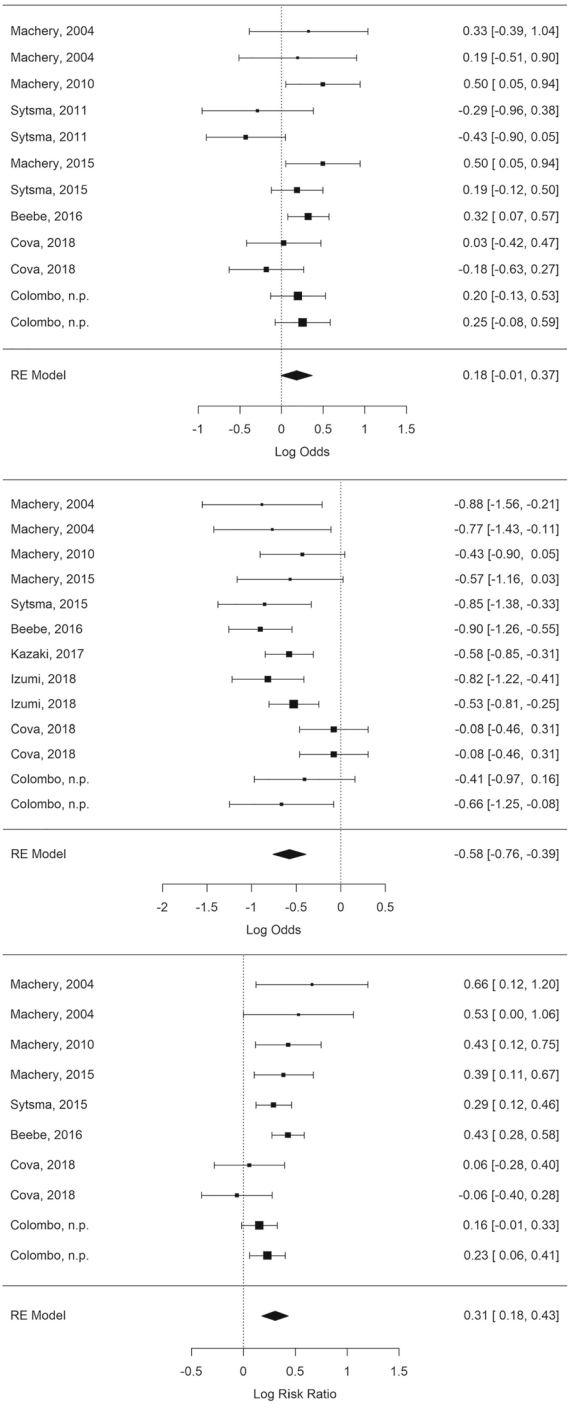

Fig. 2 Forest plot for Hypotheses 1b, 2b and 3b, respectively, showing distribution of effect sizes and confidence intervals

Hypothesis 1a The summary proportion of Causal-Historical probe responses in Westerners was 0.58 [95% CI: 0.52–0.62] with large observed heterogeneity [I^2 = 82.10%] and a prediction interval that included zero-effect [95% PI: 0.35–0.77].

Hypothesis 1b The summary proportion of Causal/Historical responses to Gödel probes in Westerners (in the original experimental design) was 0.55 [95% CI: 0.50–0.59] with moderate observed heterogeneity [I^2 = 21.89%, 95% PI: 0.35–0.77].

Hypothesis 2a The summary proportion of Causal-Historical probe responses in East Asians was 0.50 [95% CI: 0.41–0.59] with large observed heterogeneity [I^2 = 87.50%] and prediction interval [95% PI: 0.19–0.80].

Hypothesis 2b The summary proportion of Causal-Historical to Gödel probes in East Asians (in the original experimental design) was 0.36 [95% CI: 0.32–0.41] with large observed heterogeneity [I^2 = 52.59%], but a prediction interval that excluded the zero-effect value [95% PI: 0.26–0.48].

Hypothesis 3a The summary RR of Causal-Historical probe responses in Westerners versus East Asians was 1.18 [95% CI: 1.09–1.27] with moderate-to-large observed heterogeneity [I^2 = 47.42%] and a prediction interval that included zero-effect [95% PI: 0.96–1.44].

Hypothesis 3b The summary RR of Causal-Historical responses to Gödel probes in Westerners versus East Asians (in the original experimental design) was 1.36 [95% CI: 1.20–1.54] with moderate observed heterogeneity [I^2 = 30.08%] and a prediction interval that excluded the zero-effect value [95% PI: 1.03–1.80].

In addition, we wanted to make sure that these results did not depend on how we modelled the structure of the data. Most studies included in the meta-analysis contain several probes. Some studies used one sample of participants to test several probes (i.e., repeated measure design), while other studies used separate samples of participants to test each probe individually (i.e., between-subject design). Our analyses take into account the hierarchical structure of probes nested in studies, but they neglect the potential dependencies between probes in studies with repeated measures design.

To take account of these differences, we ran two additional sets of analyses. For the first analyses, we used a flat data structure (i.e., a regular RE meta-analysis). For the second analyses, we also used a regular RE meta-analysis, but included the average effect size and standard error of probe clusters with shared repeated measure design. For instance, if a study contained one sample of (Western) participants (and/or one sample of Asian participants) that responded to four probes (e.g., the two Gödel and two Jonah probes of the original study by Machery et al. (2004)), then only the average of their response was included in the analyses. The results of these analyses were similar to those of the analyses with the hierarchical data structure, thereby warranting the same conclusion (code and results of the these analyses can be found in the additional materials file).

3.3 Results of Exploratory Analyses

The width of the confidence and prediction intervals in the hypotheses tests (see Section 3.2) point to high observed inter-study variance of participants' responses. By means of a meta-analytic regression, we explored whether this variance could be explained by specific deviations in study design, such as language of the participants (see Appendix B for a list of these deviations and their descriptions). Concretely, we fitted three generalized meta-analytic regressions with the Knapp and Hartung method[10] (Knapp and Hartung 2003; Viechtbauer 2010) for the data used to test hypotheses 1a, 2a, and 3a with the identified design deviation factors as predictors and log-transformed RR (hypothesis 3a data) or Odds Ratio (hypotheses 1a and 2a data) as outcome. We fixed the sampling variance to zero, because we were only considering variance between probe outcomes. This choice is acceptable, since we are not estimating an overall effect-size weighted by the precision of each datum in the analysis. However, we did include the sample size in the model as a predictor.

For all three analyses, much of the variance was explained by the model[11]. About 47% of the variance in the probe responses by Westerners could be accounted for by deviations in study design [R^2 = 0.47, F(18,30) = 3.34, $p = 0.002$]. For probe responses by East Asians, about 64% of the variance could be accounted for [R^2 = 0.64, F(22,23) = 4.60, $p = 0.0003$]. In the comparison in probe responses between Westerners and East Asians, about 28% of the variance could be accounted for by deviations in study design [R^2 = 0.28, F(18,15) = 1.70, $p = 0.15$].[12] Although the relatively large number of predictors in relation to the size of the dataset made the results unsuited for drawing strong conclusions with certainty, a noteworthy result was that none of the design deviations, which include participant nationality and different types of probes, made a statistically significant contribution ($p < 0.05$) to all of the models.[13]

In addition, we separately analyzed studies that used an alternative formulation of the binary-response question in the Gödel probes: the "clarified narrator's perspective" e.g.,(Sytsma and Livengood 2011). Based on existing literature, this formulation was expected to be less ambiguous in eliciting participants' semantic intuitions. We found six such probes in our data set, of which three compared the causal-historical

[10]This method was used because with this method the individual coefficients are based on the t-distribution and the omnibus test statistic uses an F-distribution. In other words, this methods allows for interpretation of the result similar to a regular linear regression.

[11]Because of the fixed sampling-variance, the analysis is identical to a generalized linear model with R^2 equal to adjusted-R^2.

[12]The models do not have the same number of predictors, because of group specific variables (e.g., language of participant group) and redundant dummy predictors were left out.

[13]Specifically, for the analysis with Western participants only, significant ($p < 0.05$) predictors were 'Julius 2' probe, 'Satyricon 2' probe, 'Satyricon 3', the 'good moral valance' and 'bad moral valance' manipulations, and the 'clarified narrator's perspective' question formulation. For the East Asians only analysis, significant ($p < 0.05$) predictors were the probes 'Jonah 1', 'Jonah 2', 'Julius 1', 'Julius 2' and 'Shakespeare 1', the 'good moral valance' manipulation, and the 'sono-demonstrative' answer formulation. For the analysis that compared Western and East Asian responses, significant ($p < 0.05$) predictors were the 'Gödel 2' probe and the 'Shakespeare 1' probe. For a complete overview can be found under '2. Exploratory analyses results' in the additional materials file.

Table 2 Results of the analysis for probes with the clarified narrator's perspective proposed by (Sytsma and Livengood 2011)

Analysis	Effect Type	Effect Size	95% CI	*P* value	I^2	95% PI
Westerners	proportion	0.66	0.60–0.72	$p < 0.0001$	3.27%	0.56–0.75
East Asians	proportion	0.39	0.28–0.52	$p < 0.10$	68.10%	0.20–0.62
Comparison	RR	1.31	1.17–1.47	$p < 0.0001$	0.0%	1.17–1.47

and descriptivist responses between Westerners and East Asians; one tested East Asians only; and two tested Westerners only. To test the summary effect-size of these data, the same kind of analyses were used as for testing H1b-H3b (see Section 2.4). The results are summarized in Table 2. The proportion of causal-historical intuitions is higher for both Westerners and East Asians in the clarified narrator's perspective than in the set of all Gödel probes. In particular, for Westerners, the effect—as measured by the proportion of causal-historical responses—is more pronounced in the "clarified narrator's perspective" while it is slightly smaller for East Asians and the comparison of East Asians and Westerners. These comparisons are relative to the outcomes of the tests of H1b–H3b in Section 3.2.

Finally, we analyzed the difference in responses to non-Gödel probes. This analysis is a complement to the test of hypothesis 3b, a comparison in responses between Westerners and East Asians on all the probes except the Gödel probes. Cultural difference is absent (RR = 0.08, 95% CI =[-0.04, 0.20]), but this is not particularly surprising: the analysis consisted overwhelmingly of Jonah probes which did not show a significant difference between the samples in the original study by (Machery et al. 2004).

3.4 Evaluation of the Study's Results and Methods

Evaluation of the results of the included papers indicate absence of small-study effects and presence of evidential value, but high inter-study variance. The p-curves are right-skewed (see Figs. 3 and 4), which could be considered indicative of the set of studies containing evidential value (Simonsohn et al. 2014).[14] Figure 5 shows the funnel plots (Sterne et al. 2011) for Hypothesis 1a-3a, and Fig. 6 shows the corresponding funnel plots for Hypothesis 1b-3b. Small to large studies show symmetrical distributions around the summary effect-sizes (see Figs. 5 and 6), but for Hypothesis 1a and 2a, inter-study variance is high even among studies with high precision (i.e., large sample size, low standard error). This indicates that inter-study variance does not diminish with increasing precision as it should. No such effect is observed when cross-cultural variations are directly studied, that is, in Hypothesis 3a/b. For a quality appraisal of the methods employed in the experiments, see Appendix C.

[14]It should be noted that the reliability of p-curves is negatively related with the heterogeneity between the included results, which is quite large in the case of our meta-analyses. Also, p-curves can fail to show p-hacking and publication bias in certain scenarios where it is actually present (Carter et al. 2019; Bishop and Thompson 2016).

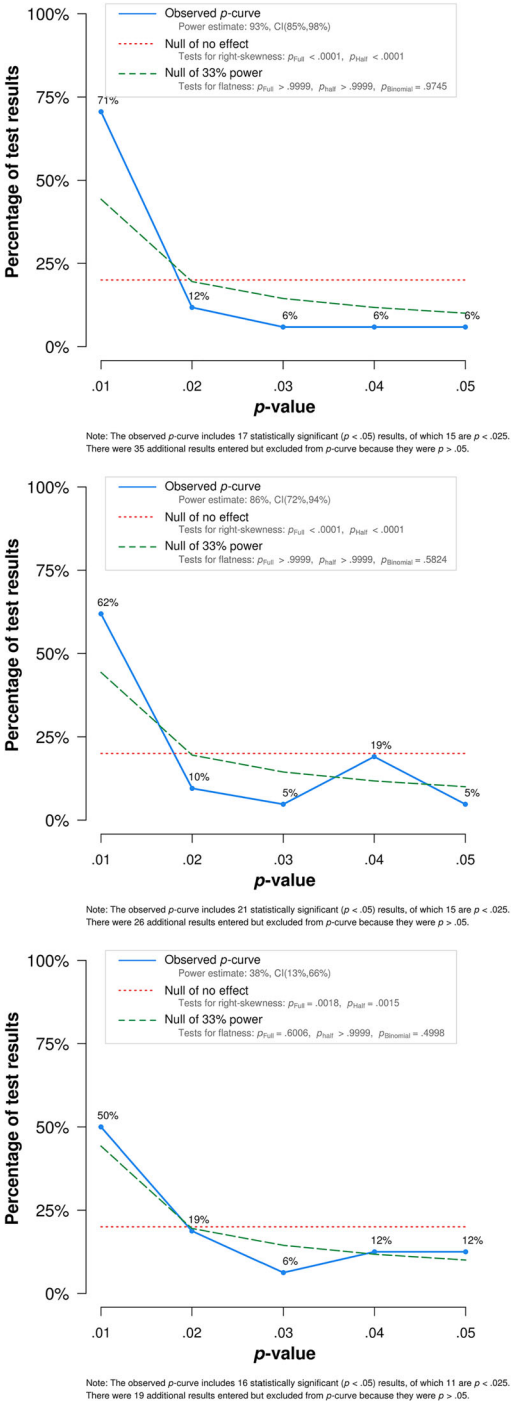

Fig. 3 P-Curve plots for Hypotheses 1a, 2a and 3a, respectively

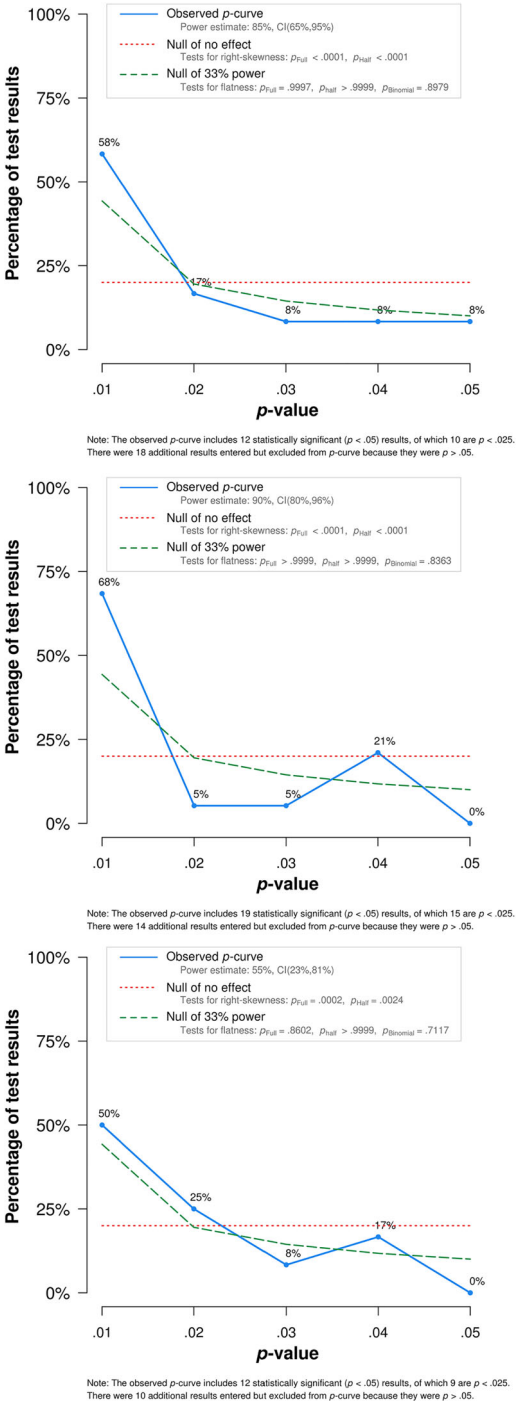

Fig. 4 P-Curve plots for Hypotheses 1b, 2b and 3b, respectively

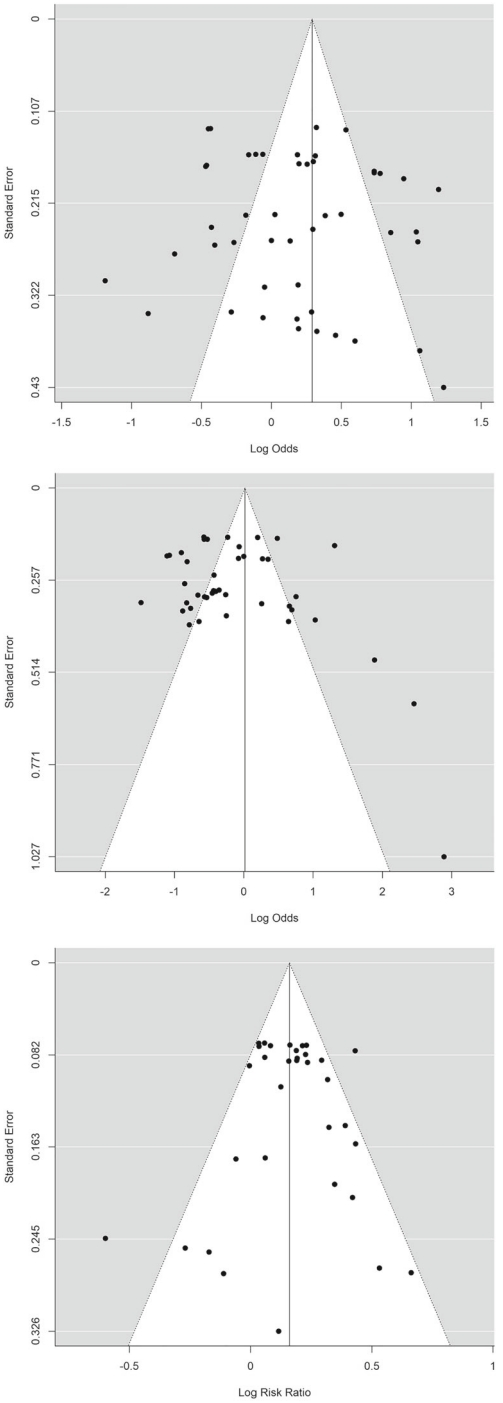

Fig. 5 Funnel plots for Hypotheses 1a (upper left figure), 2a (upper right figure) and 3a (bottom figure). x-axis: observed effect size; y-axis: precision of study as measured by standard error

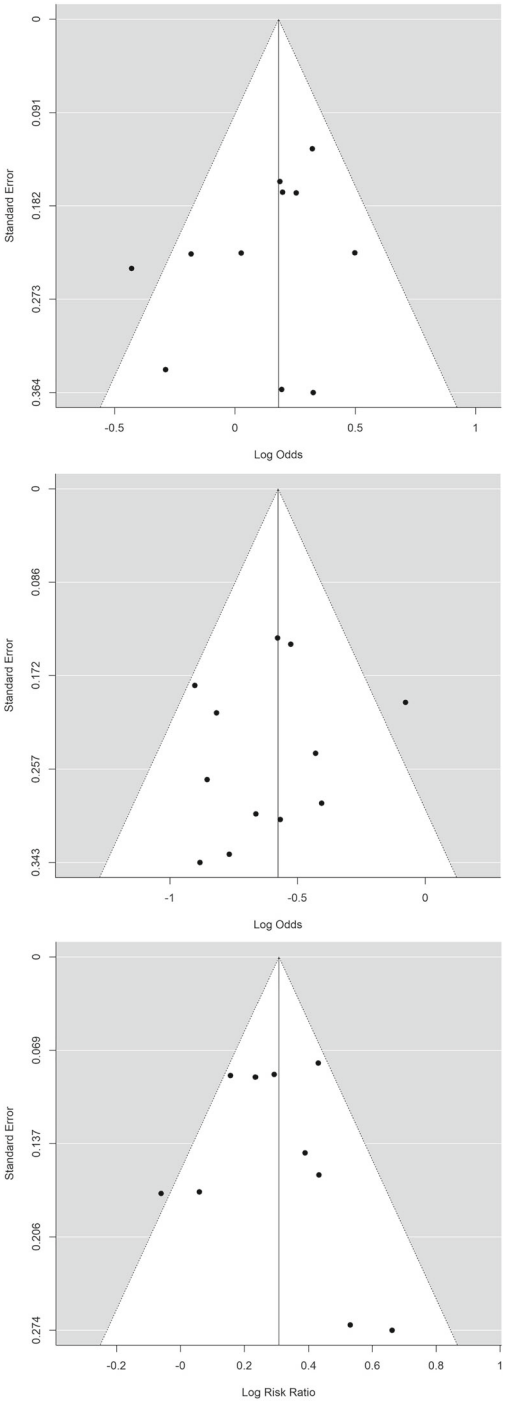

Fig. 6 Funnel plots for Hypotheses 1b (upper left figure), 2b (upper right figure) and 3b (bottom figure). x-axis: observed effect size; y-axis: precision of study as measured by standard error

4 Discussion

All in all, our meta-analysis supports the hypothesis that cross-cultural factors affect semantic intuitions about proper names (Machery et al. 2004). For four out of the six tested hypotheses, the meta-analytic confidence interval for the summary effect size does not include the zero effect value. Neither do specific analysis tools aimed at detecting publication bias or QRPs (e.g., funnel plots, *p*-curves) provide evidence of systematic suppression of negative results. The wide prediction intervals point to high inter-study variability of the data, which cannot be consistently explained by the studies' differences in experimental design.

Our study cannot test the overall scope of cross-cultural factors in eliciting intuitions about the reference of proper names, since the meta-analysis is restricted to a very specific set of probes. In addition, some aspects of the meta-analysis are limited by the methodological design of the studies and the small number of studies with respect to the differences in design between the studies. The use of a few vignettes with binary responses per study with unknown dependencies forced us to analyze the data on the level of individual probes (vignettes) and may have contributed to the large heterogeneity. However, three findings of general interest stand out.

First, there is a notable difference between the confidence intervals (CIs) for the unknown parameter and the prediction intervals (PIs) for the next observation: only two of the six PIs do not contain zero effect. The clearest difference between the CIs and the PIs is perhaps visible in the meta-analysis of Hypothesis 1a, 1b and 2a: the 95% PIs are too wide to make a theoretically meaningful prediction. The reason is that PIs take into account how the variability in the data transfers to the expected value of future data points. In a case where individual studies scatter over a wide range of points, like in the case of Hypothesis 1a, 1b, and 2a, the PIs will be considerably wider than the CIs. The same remarks apply, though with a less pronounced effect, to the other three hypotheses. For these reasons, it is not surprising that a recent replication attempt, included in (Cova et al. 2018), has not reproduced the original effect: although there is, on average, a cross-cultural effect in the predicted direction, the inter-study variance is too high to reliably predict a result in the vicinity of the meta-analytic mean. The funnel plots of Figs. 5 and 6 confirm this diagnosis: since variation in the observed proportions of causal-historical responses is extremely high, there seems to be a lot of "noise" in the data.

Second, the clarified narrator's perspective, suggested by (Sytsma and Livengood 2011), seems to push causal-historical intuitions: it increases the tendency of participants to give causal-historical responses across the board. At the same time, it reduces the cross-cultural difference between Westerners and East Asians. However, due to the small sample size for this type of probe, our finding here should be treated with caution.

Third and last, our meta-analytic regression found statistically significant dependencies of the results on variations in experimental design. However, we could not identify predictor variables that uniformly explained inter-study variance for Westerners, East Asians and the comparison of both populations. We could not decide whether this lack of consistency between the analyses is due to overfitting, or because

relevant methodological deviations from the original study were not reported in the papers, and thus function as hidden moderators. On the other hand, there are interesting dependencies on the particular probes. For example, for the population of East Asians, the tendency to give descriptivist answers is almost exclusively driven by Gödel-type probes (meta-analytic estimate of 64% as opposed to 50% overall). This is supported by the null result of the exploratory analysis of all non-Gödel probes. These results show a high sensitivity to the particular instrument for eliciting semantic intuitions, without presence of a theoretically convincing explanation (the effect is absent for Westerners). More generally, the large amount of heterogeneity and lack of uniform explanation may (partly) be due to methodological deficiencies (see Appendix C). We therefore suggest that future research on cross-cultural differences in semantic intuitions extends the range of probes in experiments, in order to avoid that substantial philosophical and psychological conclusions depend (too) much on the specifics of a particular probe (Cf., Devitt and Porot 2018).

Given these qualifications, the take-home message of our meta-analysis can be summed up as follows: there is cross-cultural variation in intuitions about the reference of proper names, but there is a high unexplained inter-study variance, compromising predictive validity; and it is not easy to disentangle this cross-cultural effect from the (random) effect of a particular study.

In relation to existing findings from other disciplines, our results are consistent with anthropologists' and linguists' work showing that there are many features of naming practices that differ across cultures and contexts of discourse (Vom Bruck and Bodenhorn 2006). However, underlying these differences, the primary pragmatic functions of person names of referring directly to a unique individual and also marking social connections, may be stable (Alford 1988; Sophia and Marmaridou 1989). Also stable may be the kinds of cognitive processes involved in understanding proper names, which probably recruit a number of structured representations (or frames) associated with the name (Valentine et al. 1996; Dancygier 2009). Given the high degree of variation, both within and across experimental groups, that our meta-analysis uncovered, it is plausible that the workings of these kinds of processes are modulated by both causal and descriptive factors in different linguistic and social contexts (Genone and Lombrozo 2012).

Due to the dependency on particular probes, we recommend that researchers interested in understanding the mechanisms underlying the processing of person names move beyond the contrast between descriptivists and causal-historical theories, and use a broader and more diverse set of vignettes in different languages and contexts. To this end, methodologies similar to those employed in cognitive psychology and neuroscience can be useful as they are designed to uncover the mechanisms and cognitive representations underlying people's semantic intuitions. For readers who are primarily interested in questions about philosophical methodology, our results confirm that philosophical scenarios are less reliable instruments for eliciting semantic intuitions than current philosophical practice seems to presume.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommonshorg/licenses/by/4.0/.

Appendix A: List of Excluded Studies

Table 3 lists all studies in our original pool that were excluded from our meta-analysis. For each study, we specify why it did not meet the inclusion criteria. Since meta-analyses are based on effect-sizes of the results of all included studies and their standard error, the same kind of effect-size must be calculable for each study in a meta-analysis. The original study used two groups, one of "Westerners" and one of "East Asians," and binary-outcomes for each probe. Hence, the relevant kind of effect-size is the Relative Risk or Odds Ratio. Based on this kind of effect-size, we excluded studies that had outcomes with more than two values and studies or samples that did not fall in the East-Asian or Western group.

Table 3 List of excluded study on semantic intuitions

APA reference	Exclusion reasons
Nichols, S., Pinillos, N. Á., & Mallon, R. (2016). Ambiguous reference. *Mind*, 125(497), 145-175.	The design did not match. The studies in this paper did not use binary-response options for their vignettes where one was the causal-historical option and the other the descriptivist option.
Domaneschi, F., Vignolo, M., & Di Paola, S. (2017). Testing the causal theory of reference. *Cognition*, 161, 1-9.	The design did not match. The studies in this paper did not use binary-response options for their vignettes where one was the causal-historical option and the other the descriptivist option.
Genone, J., & Lombrozo, T. (2012). Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology*, 25(5), 717-742.	The design did not match. The studies in this paper did not use binary-response options for their vignettes where one was the causal-historical option and the other the descriptivist option.
Grau, C., & Pury, C. L. (2014). Attitudes towards reference and replaceability. *Review of Philosophy and Psychology*, 5(2), 155-168.	The ethnicity/cultural background of the participants was mixed. The data could not be separated into Western and East Asian samples.
Islam, F. (2017) *Logical semantics in a cross-cultural perspective*. Master's thesis in English Linguistics and Language Acquisition. Supervisor: Giosué Baggio. NTNU - Trondheim, May 2017	The author did not want to share data. The Asian participants were not East-Asians, but South-Asians.
Machery, E. (2012). Expertise and intuitions about reference. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia*, 27(1), 37-54.	The ethnicity/cultural background of the participants was mixed. The data could not be separated into Western and East Asian samples.

Appendix B: List of Experimental Design Deviations

Table 4 below provides an overview of experiment design deviations from (Machery et al. 2004). We focused on several design factors. For each factor, we identified deviations from the original design, based on the information reported in the method sections of all studies in our dataset.

Appendix C: Detailed Quality Appraisal

In addition to the evaluation of the main results of the meta-analysis (see Section 3.4), we assessed the quality of the methods and measurements techniques of the included studies. Quality appraisal is a standard feature in systematic reviews [e.g., exhaustive meta-analysis of the literature on a particular topic or field of study;](Petticrew and Roberts 2008). In general, results of a methodological quality appraisal are used for

Table 4 Overview of experimental design deviation and where they are present

Factor	Values	Description	Studies that contain this deviation
Probe	Godel1, Godel2, Jonah1, Jonah2, Shakespeare1, Julius1, Julius2, Satyricon1, Satyricon2, Satyricon3	Different vignettes were used in the studies. The original study used the Gödel and Jonah vignettes. The Shakespeare, Julius, and Satyricon vignettes are about different characters and situations, though still concern reference of proper names.	Lam, 2009 (Shakespeare1, Julius1, Julius2); Machery, 2010 (Shakespeare1); Beebe, 2015 (Satyricon1, Satyricon2, Satyricon3)
Language probe Asia	English, Chinese, Japanese	In the original study, the probes were in English for the East-Asian sample. Some studies deviated by presenting the probes in Chinese or Japanese.	Lam, 2009 (Chinese); Machery, 2010 (Chinese); Machery, 2015 (Chinese); Sytsma, 2015 (Japanese); Kazaki, 2017 (Japanese); Izumi, 2018 (Japanese)
Language participant West	English, French, Dutch, Mixed	In the original study, the main language of the Western sample was English (USA based). Some studies deviated by using Westerners with a different place or origin, residence, and main language.	Machery, 2009 (French); Cova, 2018 (Dutch); Colombo, n.p. (Mixed)
Language participant asia	Chinese, Mongolian, Japanese, Mixed	In the original study, the (main) language of the East-Asian sample was Chinese (Hong Kong based). Some studies deviated by using East-Asians with a different place or origin, residence, and main language.	Machery, 2009 (Mongolian); Sytsma, 2015 (Japanese); Kazaki, 2017 (Japanese); Izumi, 2018 (Japanese); Colombo, n.p. (Mixed)

Table 4 (continued)

Factor	Values	Description	Studies that contain this deviation
Moral valence	0=original, 1=good, 2=bad	The moral valence of the vignette was altered. E.g., in the original version Gödel stole Schmidt's incompleteness theorem, but in the altered version he took credit to assist Smith.	Beebe, 2015
Perspective probe answer	0=original, 1=protagonist's perspective, 2=narrator's perspective, 3=clarified narrator's perspective	The phrasing of the questions and the binary-response options are altered. Specification is added to guide the participant in taking the perspective of the one narrating the vignette (2, 3). The 'protagonist perspective' (1) is not supposed to test semantic intuitions. These probes were excluded from the analyses.	Sytsma, 2011; Beebe, 2015 (only 3, 'clarified narrator's perspective'); Machery, 2015 (only 3, 'clarified narrator's perspective'); Sytsma, 2015; Beebe, 2016 (only 3, 'clarified narrator's perspective')
Anchoring	0=unaltered, 1=changed anchoring in the probe.	The vignette is altered in its anchoring. Specifically, in the original the protagonists learns that, for instance Gödel proved the incompleteness theorem, but further in the vignette this is said to be incorrect. In the vignettes with changed anchoring the vignette is rearranged in such a way that it starts with that Schmidt proved the incompleteness theorem.	Beebe, 2016
Answer phrasing	0=original, 'with bare noun phrases', 1='without bare noun phrases', 2=added 'sono-demonstrative', 3=added 'sono-demonstrative' and 'anaphoric predicate', 4=added 'clarifying predicate'.	The answers phrasing is changed in particular ways for the purpose of decreasing ambiguities. Values 1, 2, and 3 concern particular in Japanese translations responses. Value 4 concerns added predicate(s) that are supposed to add clarity.	Machery, 2009; Kazaki, 2017; Izumi, 2018

deciding on study inclusion/exclusion (e.g., studies below a certain quality threshold are excluded), or as predictors for assessing the robustness of the results. Since no standards of quality appraisal have yet been formulated and tested for experimental philosophy research, we moved the quality appraisal to this appendix where we indicated presence or absence of methodological features of the studies without interpreting these results.

Table 4 (continued)

Factor	Values	Description	Studies that contain this deviation
Question target	0=reference ("talking about"), 1=truth values (sentence true?)	The type of binary-response options that were used. The original used a reference style where the protagonist of the vignette is talking about the one bearing the name (causal-historical option) or the one with the descriptive property (descriptivist option). The other type is a true-or-false question, where one is the causal-historical option and the other the descriptivist option.	Machery, 2009
Asian sample	0=students from the University of Hong Kong, 1=other populations	This variable was added later, because we realized that it might be possible that the students from the University of Hong Kong might be a highly specialized group that might respond differently than those from the other university in Hong Kong, mainland China, Japan, or other parts of East-Asia.	Lam, 2009; Machery, 2009; Machery, 2010; Machery, 2015; Sytsma, 2015; Beebe, 2016; Kazaki, 2017; Cova, 2018; Izumi, 2018; Colombo, n.p.

For the quality appraisal of the studies' methods, we used parts of a validated appraisal checklist for clinical non-RCT comparative studies (Khan et al. 2001; Deeks et al. 2003). In addition, we added three items concerning generalizability from the sample to the population suggested in (Petticrew and Roberts 2008).

The checklist consisted of nine particular qualities, on which each study could be scored 'yes' if the information provided in the paper describing the study indicated it possessed this quality; 'no' if there was clear indication in the paper that the study did not possess this quality; 'unknown' if insufficient information was given in the paper about this quality; or 'NA' if the quality was not applicable to the study (e.g., group matching does not apply to a study that tests only a single group). These nine items were:

1. *Sample description*: adequacy and completeness of description of the relevant participants' characteristics. We restricted ourselves to age (mean and standard deviation) and sex (percentage of males/females) because the original study reported these characteristics and the literature does not identify other relevant characteristics.
2. *Matched groups*: similarity of participant groups in terms of characteristics that may affect the outcome (including socio-economic characteristics). In this case, the Western and East Asian participants would have to be matched on age and sex ratio.
3. *Equal group treatment*: identical treatment of participant groups. In this case, Westerners and East Asians should be tested under the same experimental conditions.

4. *Experimental description*: adequate and unambiguous description of the experiment. In this case, the studies should contain information on what kind of design was used (within-subject or between-subject), how participants were sampled, and how the participants were tested.
5. *Adequate measure*: a measurement or test is adequate when its validity and reliability are reported. This can be ensured by reference to previous measurement assessments and/or argumentation supporting its validity and reliability.
6. *Relevant measure*: a measurement or test is relevant when it is considered appropriate for testing (answering) the reported hypothesis (research question). In this case, an explanation should be given why the probe and the binary question measurement are suitable for measuring intuitions about the reference of proper names (in comparison to other potential tests or measurements).
7. *No experimenter bias*: Experimenter bias is considered absent when the report indicates that the person(s) administrating the test could not have influenced the outcome. In our case, we consider absence of experimenter bias when the administrator(s) was (were) not in the same room as the participants during the experiment.
8. *Random sampling*: a sample is considered a random sample (of the population) when all members of the population had an equal probability of participating. This is a very demanding requirement. However, it should be noted that it is an assumption of all inferential statistical tests. In our case, at least clear non-random sampling practices should be absent (e.g., convenience sampling).
9. *Representative samples*: a sample is considered representative when there are clear indications that its composition/distribution of the relevant characteristics is similar to that of the population. In our case, the samples of Western and East Asian participants should be demographically similar to the respective populations.

The methodological quality assessment was performed by two of the authors [NvD and MC]. Two of us independently scored the included studies using the quality appraisal checklist. The results of the assessment are reported in Figs. 5 and 6. None of the studies scored affirmatively on all of the methodological qualities, and in many cases there is not enough reported on the measurement and procedure of the studies to adequately infer absence or presence of a quality.[15]

Due to the low number of high-scoring studies, it was not possible to perform a sensitivity analysis where we compare the effect-size and heterogeneity of high-scoring vs. low-scoring studies. However, it was possible to run regression analyses with the qualities as predictors. Five of the nine qualities could not be included, because none scored affirmatively. The four predictors that could be included in the analyses explained between 14.0% and 34.30% of the variance between the probes (for code and results: see the additional materials file).

[15] The original statistical analysis by (Machery et al. 2004) is flawed on several levels, most egregiously in the aggregation of binary scores across probes and the use of the *t*-test for comparing the sample means across populations.

Table 5 Assessment of methods and measurement qualities 1 through 4

First author	Year	Study No.	Sample description	Matched groups	Equal group treatment	Experimental description
Machery	2004	1	No	No	Yes	Yes
Lam	2009	1	No	No	Yes	Yes
		2	No	No	Yes	Yes
Machery	2009	1	Yes	No	Unknown	No
Machery	2010	1	No	Unknown	Unknown	No
		2	Yes	No	NA	Yes
Sytsma	2011	1	Yes	NA	NA	Yes
		2	Yes	NA	NA	Yes
		2	Yes	NA	NA	Yes
		3	Yes	NA	NA	Yes
Beebe	2015	1	Yes	No	Yes	Yes
		2	Yes	NA	NA	Yes
		3	Yes	NA	NA	Yes
		4	Yes	NA	NA	Yes
Machery	2015	1	Yes	No	Unknown	Yes
		2	Yes	No	Unknown	Yes
Machery	2015	3	Yes	NA	NA	Yes
Sytsma	2015	1	Yes	No	No	Yes
Beebe	2016	1	Yes	No	Yes	Yes
		2	Yes	No	Yes	Yes
Izumi	2017	1	Yes	NA	NA	Yes
		2	Yes	NA	NA	Yes
Kazaki	2017	1	No	NA	NA	Yes
Cova	2018	1	Yes	No	No	Yes
Colombo	n.p.	1	NA	Unknown	Yes	NA

Table 6 Assessment of methods and measurement qualities 6 through 9

First Author	Publication Year	Study Number	Adequate Measurement	Relevant Measurement	No experimenter bias	Random Sampling	Representative Sample(s)
Machery	2004	1	No	Unknown	Unknown	Unknown	No
Lam	2009	1	No	Unknown	Unknown	Unknown	No
		2	No	Unknown	Unknown	Unknown	No
Machery	2009	1	No	Unknown	Unknown	Unknown	Unknown
Machery	2010	1	No	Unknown	Unknown	Unknown	No
		2	No	Unknown	Unknown	Unknown	No
Sytsma	2011	1	No	Unknown	Unknown	Unknown	No
		2	No	Unknown	Yes	Unknown	Unknown
		3	No	Unknown	Yes	Unknown	No
		4	No	Unknown	No	No	Unknown
Beebe	2015	1	No	Unknown	Yes	Unknown	No
		2	No	Unknown	Unknown	Unknown	No
		3	No	Unknown	Yes	Unknown	No
		4	No	Unknown	Yes	Unknown	No
Machery	2015	1	No	Unknown	Unknown	Unknown	No
		2	No	Unknown	Unknown	Unknown	No
Machery	2015	1	No	Unknown	Unknown	Unknown	No
Sytsma	2015	1	No	Unknown	Unknown	Unknown	No
Beebe	2016	1	No	Unknown	Yes	Unknown	No
		2	No	Unknown	Yes	Unknown	No
Izumi	2017	1	No	Unknown	Unknown	Unknown	No
		2	No	Unknown	Unknown	Unknown	No
Kazaki	2017	1	No	Unknown	Yes	Unknown	Unknown
Cova	2018	1	No	Unknown	No	Unknown	No
Colombo	n.p.	1	No	Unknown	Yes	Unknown	No

In an effort to improve the replicability, and cross-disciplinary comparability of this line of research on semantic intutions, we encourage researchers to consider quality assessment schemes from other disciplines (e.g., in medicine), in addition to other checks on research quality, like high-powered replication studies e.g.,(Cova et al. 2018), checking for statistical reporting errors e.g.,(Colombo et al. 2018), and metascience tools for the uncovering publication bias and questionable research practices e.g.,(Stuart et al. 2018).

Acknowledgments We would like to thank Joshua Knobe, Edouard Machery, and Justin Sytsma for their comments and criticisms on previous versions of this paper. We are also grateful to Phoebe Mui for helping us with data collection in Hong Kong and China. This work was financially supported by the Alexander von Humboldt Foundation [M. C.].

References

Alford, R.D. 1988. Naming and identity: A cross-cultural study of personal naming practices. HRA Flex Books.

Bishop, D.V., and P.A. Thompson. 2016. Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ* 4: e1715.

Bright, W. 2003. What is a name? reflections on onomastics. *Language and Linguistics* 4(4): 669–681.

Carter, E.C., F.D. Schönbrodt, W.M. Gervais, and J. Hilgard. 2019. Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science* 2(2): 115–144.

Colombo, M., G. Duev, M.B. Nuijten, and J. Sprenger. 2018. Statistical reporting inconsistencies in experimental philosophy. *PloS one* 13(4): e0194360.

Cova, F., B. Strickland, A. Abatista, A. Allard, J. Andow, M. Attie, J. Beebe, R. Berninas, J. Boudesseul, M. Colombo, F. Cushman, R. Diaz, N.N.N. van Dongen, V. Dranseika, B.D. Earp, A.G. Torres, I. Hannikainen, J.V. Hernández-conde, W. Hu, F. Jaquet, K. Khalifa, H. Kim, M. Kneer, J. Knobe, M. Kurthy, A. Lantian, S. Liao, E. Machery, T. Moerenhout, C. Mott, M. Phelan, J. Phillips, N. Rambharose, K. Reuter, F. Romero, P. Sousa, J. Sprenger, E. Thalabard, K. Tobia, H. Viciana, D. Wilkenfeld, and X. Zhou. 2018. Estimating the reproducibility of experimental philosophy Review of Philosophy and Psychology 1–36.

Dancygier, B. 2009. Genitives and proper names in constructional blends New directions in cognitive linguistics 161–184.

Dancygier, B., and L. Vandelanotte. 2017. Viewpoint phenomena in multimodal communication. *Cognitive Linguistics* 28(3): 371–380.

Deeks, J., J. Dinnes, R. D'amico, A. Sowden, C. Sakarovitch, F. Song, M. Petticrew, and D. Altman. 2003. Evaluating non-randomised intervention studies Health technology assessment Winchester, England 7 27 iii–x.

Devitt, M., and N. Porot. 2018. The reference of proper names: Testing usage and intuitions. Cognitive Science. http://doi.org/10.1111/cogs.12609.

Frege, G. 1892. Über Sinn und Bedeutung Zeitschrift für Philosophie und philosophische Kritik 100:25–50.

Genone, J., and T. Lombrozo. 2012. Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology* 25(5): 717–742.

Higgins, J.P., and S. Green. 2011. Cochrane handbook for systematic reviews of interventions, volume 4. John Wiley & Sons.

Higgins, J.P., S.G. Thompson, J.J. Deeks, and D.G. Altman. 2003. Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal* 327(7414): 557.

Higgins, J.P., S.G. Thompson, and D.J. Spiegelhalter. 2009. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A Statistics in Society* 172(1): 137–159.

Ioannidis, J.P., N.A. Patsopoulos, and E. Evangelou. 2007. Uncertainty in heterogeneity estimates in meta-analyses. *Bmj* 335(7626): 914–916.

Izumi, Y., M. Kasaki, Y. Zhou, and S. Oda. 2018. Definite descriptions and the alleged east–west variation in judgments about reference. *Philosophical Studies* 175(5): 1183–1205.

Khan, K.S., G. Ter Riet, J. Glanville, A.J. Sowden, J. Kleijnen, and et al. 2001. Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews 4 2n. NHS Centre for Reviews and Dissemination.

Knapp, G., and J. Hartung. 2003. Improved tests for a random effects meta-regression with a single covariate. *Statistics in medicine* 22(17): 2693–2710.

Kripke, S. 1980. Naming and necessity Harvard University Press Cambridge/MA.

Lam, B. 2010. Are cantonese speakers really descriptivists? revisiting Cross-Cultural semantics. *Cognition* 115: 320–332.

Machery, E. 2017. Philosophy within its proper bounds Oxford University Press Oxford.

Machery, E., R. Mallon, S. Nichols, and S. Stich. 2004. Semantics Cross-Cultural style. *Cognition* 92: 1–12.

Machery, E., C.Y. Olivola, and M. LeBlanc. 2009. Linguistic and metalinguistic intuitions in the philosophy of language. *Analysis* 69: 689–694.

Machery, E., J. Sytsma, and M. Deutsch. 2015. Speaker's reference and cross-cultural semantics. In: Bianchi, A., editor, On Reference, pages 62–76, Oxford University Press Oxford.

Marmaridou, S.S. 2000. Pragmatic meaning and cognition, volume 72. John Benjamins Publishing.

Müller, H.M. 2010. Neurolinguistic findings on the language lexicon: the special role of proper names Chinese Journal of Physiology 53 6.

Nisbett, R., K. Peng, I. Choi, and A. Norenzayan. 2001. Culture and systems of thought: Holistic versus analytic cognition. *Psychological review* 108: 291–310.

O'Mara, A.J. 2008. Methodological and substantive applications of meta-analysis: Multilevel modelling, simulation, and the construct validation of self-concept Unpublished doctoral dissertation Oxford University.

Petticrew, M., and H. Roberts. 2008. Systematic reviews in the social sciences a practical guide Blackwell Publishing Malden.

Proverbio, A.M., S. Mariani, A. Zani, and R. Adorni. 2009. How are 'barack obama' and 'president elect' differentially stored in the brain? an erp investigation on the processing of proper and common noun pairs. *PloS one* 4(9): e7126.

Raudenbush, S.W. 2009. Analzing effect sizes: Random effects models. In: Cooper, H., V, H. L., and C, V. J., editors, The handbook of research synthesis and meta-analysis, pages 295–315. Russell Sage Foundation, New York.

Riley, R.D., J.P. Higgins, and J.J. Deeks. 2011. Interpretation of random effects meta-analyses. *Bmj* d549: 342.

Russell, B. 1905. On Denoting. *Mind* 14: 479–493.

Searle, J.R. 1958. Proper names. *Mind* 266: 166–173.

Semenza, C. 2006. Retrieval pathways for common and proper names. *Cortex* 42(6): 884–891.

Simonsohn, U., L.D. Nelson, and J.P. Simmons. 2014. P-curve: a key to the file-drawer. *Journal of Experimental Psychology:, General* 143(2): 534.

Sophia, A., and S. Marmaridou. 1989. Proper names in communication. *Journal of linguistics* 25(2): 355–372.

Sterne, J.A., A.J. Sutton, J.P. Ioannidis, N. Terrin, D.R. Jones, J. Lau, J. Carpenter, G. Rücker, R.M. Harbord, C.H. Schmid, and et al. 2011. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *Bmj* d4002: 343.

Stuart, M., D. Colaço, and E. Machery. 2018. P-curving x-phi: Does experimental philosophy have evidential value?. https://doi.org/10.31234/osf.io/p7ube.

Sytsma, J., and J. Livengood. 2011. A new perspective concerning experiments on semantic intuitions. *Australasian Journal of Philosophy* 89: 315–332.

Valentine, T., T. Brennen, and S. Brédart. 1996. On the importance of being ernest: the cognitive psychology of proper names.

Viechtbauer, W. 2005. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* 30(3): 261–293.

Viechtbauer, W. 2010. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* 36(3): 1–48.

Vom Bruck, G., and B. Bodenhorn. 2006. The anthropology of names and naming. Cambridge University Press Cambridge.

Wang, L., R.G. Verdonschot, and Y. Yang. 2016. The processing difference between person names and common nouns in sentence contexts: an erp study. *Psychological research* 80(1): 94–108.
Yen, H.-L. 2006. Processing of proper names in mandarin chinese: A behavioral and neuroimaging study.

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