Forthcoming in the *Australasian Journal of Philosophy*. Please cite published version when one is available.

Semantics, Cross-Category Style

Jeske Toorman and Jussi Haukioja

NTNU Trondheim

Abstract: Ever since Machery *et al.* first decided to test whether non-philosophers assign reference in accordance with the causal-historical account, the reference of proper names has been tested by means of setups modelled on Kripke's Gödel and Jonah cases. Over the years, the use of these setups as a means to test theories of reference has attracted much criticism. However, previous follow-up studies have supposedly accounted for these criticisms, for the most part without changing the original outcome. We conducted experiments suggesting that participants' responses in these setups never tracked what they are supposed to track. In our study, we tested the setup itself by using analogues of different setups modelled on Kripke's Gödel and Jonah cases. Instead of proper names, our analogues featured terms for which we have independent reason to believe that the causal-historical account is not true. The analogues elicited large proportions of supposed causal-historical responses.

Keywords: causal theory of reference, descriptivism, categorization, experimental semantics

1. Introduction and Background

For the past two decades, the reference of proper names has been tested by means of setups modelled on Kripke's Gödel and Jonah cases (for example Machery, Mallon, *et al.* 2004; Domaneschi, Vignolo and Di Paola 2017; Li, Liu, *et al.* 2018; Devitt and Porot 2018). In fact, all existing experimental work on the reference of proper names has been done using setups following their basic structure.

Kripke (1980) presented a thought experiment in which, rather than Gödel, another man named 'Schmidt' proved the incompleteness of arithmetic. Gödel merely stole the proof and ended up taking the credit for it. According to classical descriptivism, the reference of proper names like 'Gödel' is determined by associated descriptions. Supposing that many people have only one belief about Gödel, namely that he proved the incompleteness of arithmetic, descriptivism entails that in this case, these people would be talking about the person who actually proved the theorem. According to the causal-historical account, by contrast, a proper name refers to the person who was first given that name in an initial act of baptism, which is at the source of a communicative or causal chain by virtue of which reference is preserved.

According to the causal-historical account, people using ‘Gödel’ and who are part of this chain are thus talking about the person who stole the proof.

When first presented with setups modelled on this case, about two thirds of American and one third of Chinese participants gave the purported causal-historical answer (Machery, Mallon, *et al.* 2004), thereby suggesting that there is cross-cultural as well as intra-cultural variation in how individuals assign reference.

The kind of setup at issue has been criticized on many grounds (see, for example, Ludwig 2007; Deutsch 2009; Martí 2009; Sytsma and Livengood 2011). By now, however, these criticisms have supposedly been accounted for. Nevertheless, with the exception of Devitt and Porot’s (2018) study, the results seem to replicate (see, for example, Machery, Sytsma, *et al.* 2015; Machery, Olivola, and De Blanc 2009; Sytsma, Livengood, *et al.* 2015)[1]. Further, Li, Liu, *et al.* (2018) claim to show that the cross-cultural differences are already present in children. It seems, then, that there really are cross-cultural and intra-cultural differences of a sort that are of relevance to theories of reference.

We suspect that none of the thus far mentioned criticisms get to the real problem with setups modelled on Kripke’s Gödel and Jonah cases, and that responses to these setups track something other than whether participants assign reference in accordance with the causal-historical account or descriptivism. We have two worries. First, the vignettes used explicitly state that Gödel stole the proof of the incompleteness theorem, or make an analogous claim. The worry then is that subjects merely repeat (or paraphrase) the narrator, when indicating that they take ‘Gödel’ to refer to the person who stole the proof (the purported causal-historical response), without considering who or what the name refers to. Second, because the participants are told (something analogous to) that Gödel stole the proof, a descriptivist about names can also account for the purported causal-historical response by claiming that the subjects associate the description ‘the person who stole the proof’ with the name. If so, and if subjects also base their answer on what they take ‘Gödel’ to refer to in their own language, descriptivism and the causal-historical account make the same prediction concerning participants’ responses. Hence, it is not clear to us that the purported causal-historical responses should really be taken to count in favour of the causal-historical account.

We should note in advance that our concern in this paper is not with the question of whether these setups can show us anything about the reliability of *philosophers’* responses to

[1] For a recent overview of successful and failed replication attempts, see Machery 2024; for a meta-analysis, see van Dongen, Colombo, *et al.* 2021.

thought experiments, such as Kripke's Gödel case. Rather, our concern is with the relevance of these setups to the question of which theory of reference is true.[2]

In order to find out whether these setups track what they are supposed to track, we created vignettes and probe questions that serve as a control. Instead of proper names, our vignettes and probe questions used terms for games and terms for tools as the target term. Apart from this difference, they were as similar as possible in structure and content to the original vignettes and probe questions that have been used to test the reference of proper names. It is at least *prima facie* plausible to think that the causal-historical theory is not true of terms for tools and games, and that the extensions of these terms are rather determined by intended functions and game rules, respectively.

Our aim here is *not* to test theories of reference, but the experimental setup itself, by investigating whether supposed causal-historical responses can be elicited with terms of which the causal-historical account is not true. If these setups track what they are supposed to track, our controls should *not* elicit a significant proportion of causal-historical responses, since they target terms of which the causal-historical account is false. If they were to do so, they cannot serve as evidence for a causal-historical account, and purported causal-historical responses to the original setups, featuring proper names, cannot count as evidence for the causal-historical theory, either.

2. Experiment 1

2.1. Materials and Methods

We created vignettes and probe questions that are analogous to the vignettes and probe questions used by Machery, Mallon, *et al.* (2004), Li, Liu, *et al.* (2018), and Devitt and Porot (2018). From Machery, Mallon, *et al.* and Li, Liu, *et al.*, we used one vignette and corresponding probe question each, namely the Gödel and the Super Dog Race vignette and probe question, respectively. Both were modelled on Kripke's Gödel case. In the case of Li, Liu, *et al.*'s Super Dog Race vignette, we made analogues of the shortened version mentioned by the authors in their 2018 paper, instead of the longer version they used in their experiment.[3] From Devitt and Porot's study, we used two vignettes, one modelled on Kripke's Gödel case, featuring 'Tsu Ch'ung Chih' and one modelled on Kripke's Jonah case, featuring 'Ambriorix.' We created analogues of their two kinds of truth-value judgment tasks, one where a judgement

[2] Although Machery, Mallon, *et al.* framed their 2004 paper as having the former concern, in later studies Machery (e.g., Mallon, Machery, *et al.* 2009; Machery 2011) takes his results also to bear on the latter question. Subsequent studies, such as Devitt and Porot 2018, are more clearly framed as having the latter concern.

[3] The same shortened version of the vignette (with a different probe question) has also been used by Domaneschi and Vignolo (2020).

to the effect that the statement is true is in accordance with descriptivism (TVJ-D), and one where a judgement to the effect that the statement is true is in accordance with the causal-historical theory (TVJ-CH). We made two different analogues of each condition, one featuring a term for a tool and one featuring a term for a game. In total, we thus had twelve different conditions. In all conditions, we intentionally did not capitalize the target term, thereby discouraging a reading according to which the game or tool at issue is trademarked.

As an example, Devitt and Porot used the following vignette and probe question in their 'Tsu Ch'ung Chih' TVJ-CH condition:

> Students in astronomy classes in Hong Kong are told that a man called 'Tsu Ch'ung Chih' first determined the precise time of the summer and winter solstices. This is the only thing that typical Hong Kongers ever hear about this man. Now suppose that that man did not make the discovery he is credited with. He stole it from an astronomer who died soon after making the discovery. But the theft remained entirely undetected and so the man that Hong Kongers have been told about became famous for the discovery of the precise times of the solstices.
>
> TVJ-CH: Having read the above story and accepting that it is true, please indicate below whether you think the following statement is true or false. Tsu Chu'ung Chih was a thief and a liar. True / False

Our analogue of this case featuring a term for a tool, namely 'magnometer,' was as follows:

> Students in astronomy classes in Hong Kong are told that an instrument called 'magnometer' is used to measure the strength and direction of magnetic fields. This is the only thing typical Hong Kongers ever heard about this instrument. Now suppose that these instruments cannot be used to measure the strength and direction of magnetic fields. They are instruments for measuring the spectrum of light. But the fact that they don't have this function remained entirely undetected and so the instrument the Hong Kongers have been told about became famous for its use in measuring the strength and direction of magnetic fields.
>
> TVJ-CH: Having read the above story and accepting that it is true, please indicate below whether you think the following statement is true or false. A magnometer is an instrument that can be used to measure the spectrum of light. True / False

To repeat, our aim is only to test the setup itself, not theories of reference for game and tool terms. As such, any potential flaws in the original vignettes and probe questions are intentionally carried over to our analogues, to test whether they work as intended.

We recruited 105 British participants via the online platform Prolific. All participants were native English speakers. After answering mandatory background questions, each participant was randomly assigned to one attention check question and four of the twelve conditions. No participant saw more than one condition containing an analogue of the same vignette. The conditions were counterbalanced for order, such that of all participants receiving the same four vignettes, half of them received them in one order, and the other half in the reverse. We divided the conditions such that the probe questions analogous to those of Devitt and Porot received half the amount of responses compared to the other conditions. This was done because Devitt and Porot elicited a higher proportion of supposed causal-historical responses, making it easier to detect a relatively low proportion of supposed causal-historical responses to our analogues, as compared to those received by Devitt and Porot. Additionally, we take the weight of the evidence to be carried by the responses to the TVJ-D and TVJ-CH conditions of the same vignette together. All materials created for the studies reported in this paper, as well as all the responses, can be accessed at: https://osf.io/mqjbs/?view_only=4cf6f394dc7940608390521fb680b7ff.

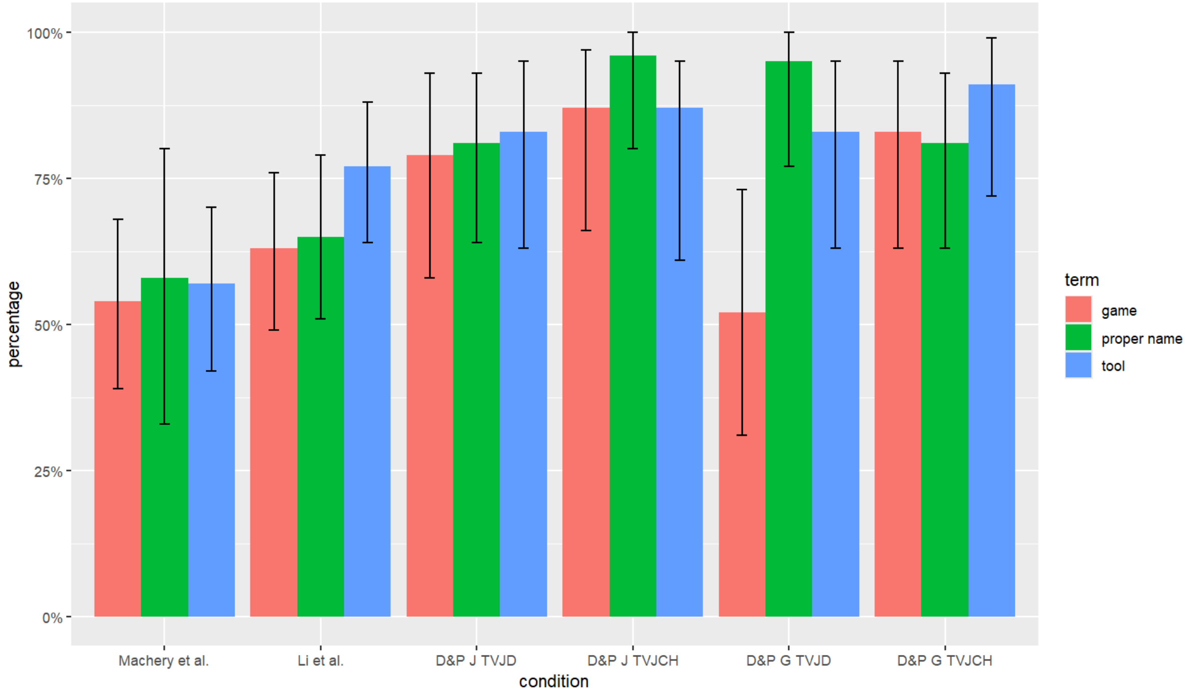
2.2. Results

Of the 105 participants, 75 identified themselves as female, 29 as male, and one as neither male nor female. Their average age was 38.29 (SD: 12.14). Eleven subjects failed the attention check question. Their results have been omitted from further analyses only when this was done in the analysis of the results of the original vignettes, to allow for a fairer comparison between responses to the original cases and to our analogues. As all tasks indicating accordance with the causal-historical account contained only two response options, we compared these responses to all conditions individually, including the originals, against chance (that is, 50% causal-historical responses) by means of a one-sample, two-sided exact binomial test, to exclude the possibility that the proportion of causal-historical responses was due to chance. Additionally, we compared the responses to the analogues with the responses (by adult Westerners) to the original conditions by means of a two-sided Fisher's exact test. The results are summarized in Figure 1 and Table 1.

In all cases but one, our analogues elicited roughly the same percentage of causal-historical responses as the originals. The difference reached statistical significance ($p < 0.05$) only for one of the game term analogues of Devitt and Porot's 'Tsu Ch'ung Chih' vignette. [4] Now the fact that we did not find a statistically significant difference between the responses to our analogues and the responses to the original conditions does not imply that there is no difference between the groups. However, the confidence intervals do suggest that if there are differences, these differences cannot be very large. Assuming the causal-historical account is indeed false of the target terms used in our analogues, that at minimum suggests that a relatively large proportion of responses to the original cases do not track what they are supposed to track, and at most that none of them do.

[4] Using the Bonferroni adjusted alpha level of 0.004 (0.05/12) still indicates that there is a statistically significant difference between the results of Devitt and Porot's original 'Tsu Ch'ung Chih' TVJ-D condition and our analogue featuring a game term.

Figure 1: *Percentage of causal-historical answers in Experiment 1*



Note: 'D&P J' stands for Devitt and Porot's 'Ambiorix' vignette modelled on Kripke's Jonah case; 'D&P G' stands for Devitt and Porot's 'Tsu Ch'ung Chih' vignette, modelled on Kripke's Gödel case. The black lines in the middle of the bar represent the 95% confidence interval of the percentage of causal-historical answers.

Table 1: *Results of the original vignettes and their analogues in Experiment 1.*

Original vignettes					
Vignette:	Percentage causal-historical answers:	95% CI:	Independence from chance; p-value:		N:
Machery et al.	58	33-80	0.65		19
Li et al.	65	51-79	0.04		47
D&P ‘Ambiorix’ TVJ-D	81	64-93	<0.001		32
D&P ‘Ambiorix’ TVJ-CH	96	80-100	<0.001		26
D&P ‘Tsu Ch’ing Chih’ TVJ-D	95	77-100	<0.001		22
D&P ‘Tsu Ch’ing Chih’ TVJ-CH	81	63-93	<0.001		31
Tool analogues					
Vignette:	Percentage causal-historical answers:	95% CI:	Independence from chance; p-value:	Comparison; p-value:	N:
Machery et al.	57	42-70	0.41	1	53
Li et al.	77	64-88	<0.001	0.27	53
D&P ‘Ambiorix’ TVJ-D	83	63-95	0.002	1	24
D&P ‘Ambiorix’ TVJ-CH	87	61-95	0.003	0.17	23
D&P ‘Tsu Ch’ung Chih’ TVJ-D	83	63-95	0.002	0.35	24
D&P ‘Tsu Ch’ung Chih’ TVJ-CH	91	72-99	<0.001	0.44	23
Game analogues					
Vignette:	Percentage causal-historical answers:	95% CI:	Independence from chance; p-value:	Comparison; p-value:	N:
Machery et al.	54	39-68	0.68	0.79	52
Li et al.	63	49-76	0.07	0.44	52
D&P ‘Ambiorix’ TVJ-D	79	58-93	0.006	1	24
D&P ‘Ambiorix’ TVJ-CH	87	66-97	<0.001	0.33	23
D&P ‘Tsu Ch’ung Chih’ TVJ-D	52	31-73	1	<0.001	23
D&P ‘Tsu Ch’ung Chih’ TVJ-CH	83	63-95	0.002	1	24

2.3. Discussion

Experiment 1 shows that vignettes modelled on Kripke’s Gödel and Jonah cases, but featuring terms for games and tools, elicit large proportions of purportedly causal-historical responses. The terms used in the analogues were not trademarked names for tools and games, but rather general terms. If we are correct in thinking that the causal-historical account is not true of these terms, our results show that Gödel and Jonah-type cases elicit significant proportions of purportedly causal-historical responses also for terms of which the causal-historical theory is false. This, in turn, would show that most or all test subjects’ responses to such cases tell us nothing about whether the causal-historical theory is true of *any* term, proper names included.

One might object that reference assignments have been found to vary cross-culturally (for example, Machery, Mallon, *et al.* 2004; Machery, Olivola, *et al.* 2023). At the same time, the large proportion of supposedly causal-historical responses to our analogues is only a problem for Gödel and Jonah-type cases if these responses are from participants with the same cultural background as the participants from the original studies. After all, if it were to turn out, for example, that participants with the same cultural background as the participants responding to our analogues were to give the purported descriptivist response to the original vignettes and probe questions featuring proper names, there *would* be a contrast between proper names on the one hand, and terms for tools and games on the other, after all, and a more plausible conclusion would be that participants with said cultural background just turn out to assign referents to proper names in accordance with descriptivism and to terms for tools and games in accordance with the causal-historical account. However, our participants consisted exclusively of British participants, whereas the experiments of which we made analogues consisted exclusively of American participants. Given the possibility of cross-cultural variation in reference assignments, we cannot take for granted that the responses of British participants to the original vignettes are in line with the responses of the American test subjects who participated in the original studies.

Experiment 2 was conducted to test whether we can similarly elicit relatively large proportions of supposed causal-historical answers both with our analogues as well as with the originals, with participants from the same cultural background. Additionally, we wanted to replicate Experiment 1 with each participant receiving only one vignette and probe question instead of several, such as to remove any doubts about possible interference caused by previously read vignettes and probe questions.

3. Experiment 2

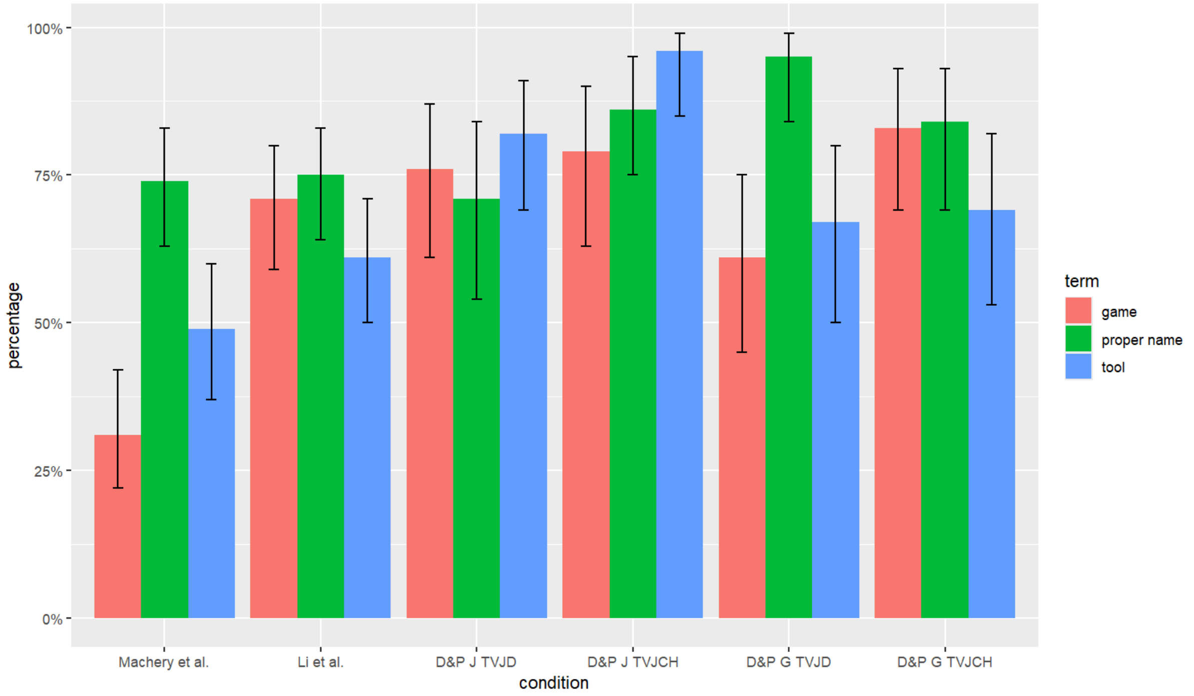
3.1. Materials and Methods

In Experiment 2, we presented a group of British participants with exactly one vignette and corresponding probe question each. The set of vignettes and probe questions included, on the one hand, ones identical to those used in the original studies on proper names and, on the other, analogues identical to those we used in Experiment 1. In the case of the original vignette and probe question featuring a proper name used by Li, Liu, *et al.*, we used the shortened version of which we also made analogues. Half of the participants were randomly assigned to the original vignette by Machery, Mallon, *et al.*, the original vignette by Li, Liu, *et al.*, or one of their analogues featuring a tool or game term. The remaining participants were randomly assigned to either one of the original vignettes by Devitt and Porot, or one of their analogues. The (analogues of the) probe questions from Devitt and Porot received fewer participants than the other conditions for the same reason as in Experiment 1. After providing their informed consent and answering mandatory background questions, all participants were given one attention check question, one vignette, and one corresponding probe question.

3.2. Results

1193 British participants were recruited via the crowdsourcing site Prolific. [5] All participants were native English speakers. Their mean age was 41.98 years (SD: 13.31). 67.41% identified as female, 31.55% as male, 0.78% as ‘other’, and 0.26% preferred not to disclose this information. Of all participants, 150 failed the attention check question. This time, we omitted their results from further analysis in all cases. We again compared all responses against chance by means of a one-sample, two-sided exact binomial test. Additionally, we compared the results of the analogues to our results of the original vignettes and probe questions featuring proper names, and we compared our results of the original vignettes and probe questions to the results reported in the original studies by means of a two-sided Fisher’s exact test. The results are displayed in Figure 2 and Table 2.

Figure 2 :*Percentage of causal-historical answers in Experiment 2*



Note: ‘D&P J’ stands for Devitt and Porot’s ‘Ambiorix’ vignette modelled on Kripke’s Jonah case; ‘D&P G’ stands for Devitt and Porot’s ‘Tsu Ch’ung Chih’ vignette, modelled on Kripke’s Gödel case. The black lines in the middle of the bar represent the 95% confidence interval of the percentage of causal-historical answers.

[5] The chosen sample size would give a confidence level of 95% that the real value is within 12% of the measured value, on the assumption that the proportion of supposed causal-historical responses in all cases is the same as those reported in the original studies, and that no more than 15% of the participants will fail the attention check question.

Table 2: *Results of the original vignettes and their analogues in Experiment 2*

Replication original vignettes					
Vignette:	Percentage causal-historical answers:	95% CI:	Independence from chance; p-value:	Comparison; p-value:	N:
Machery et al.	74	63-83	<0.001	0.171	85
Li et al.	75	64-83	<0.001	0.323	90
D&P ‘Ambiorix’ TVJ-D	71	54-84	0.012	0.414	41
D&P ‘Ambiorix’ TVJ-CH	86	75-95	<0.001	0.410	49
D&P ‘Tsu Ch’ing Chih’ TVJ-D	95	84-99	<0.001	1	42
D&P ‘Tsu Ch’ing Chih’ TVJ-CH	84	69-93	<0.001	0.765	43
Tool analogues					
Vignette:	Percentage causal-historical answers:	95% CI:	Independence from chance; p-value:	Comparison; p-value:	N:
Machery et al.	49	37-60	0.911	0.001	80
Li et al.	61	50-71	0.044	0.077	87
D&P ‘Ambiorix’ TVJ-D	82	69-91	<0.001	0.222	50
D&P ‘Ambiorix’ TVJ-CH	96	85-99	<0.001	0.270	46
D&P ‘Tsu Ch’ung Chih’ TVJ-D	67	50-80	0.044	0.002	42
D&P ‘Tsu Ch’ung Chih’ TVJ-CH	69	53-82	0.020	0.131	42
Game analogues					
Vignette:	Percentage causal-historical answers:	95% CI:	Independence from chance; p-value:	Comparison; p-value:	N:
Machery et al.	31	22-42	<0.001	<0.001	96
Li et al.	71	59-80	<0.001	0.606	78
D&P ‘Ambiorix’ TVJ-D	76	61-87	<0.001	1	46
D&P ‘Ambiorix’ TVJ-CH	79	63-90	<0.001	0.545	38
D&P ‘Tsu Ch’ung Chih’ TVJ-D	61	45-75	0.184	<0.001	46
D&P ‘Tsu Ch’ung Chih’ TVJ-CH	83	69-93	<0.001	1	42

3.3 Discussion

In Experiment 2, British participants gave predominantly causal-historical answers in response to the original vignettes and probe questions featuring proper names, just as American test subjects did in the original studies. In no case did the difference between the results reported in the original papers and the results obtained by us reach statistical significance. Together with the results of our analogues in Experiment 2, this shows that we do get predominantly causal-historical answers for both proper names and terms for tools and games, in otherwise identical setups, from test subjects coming from the same cultural background and in otherwise equal conditions.

We continued to get problematically high proportions of supposedly causal-historical answers for tool terms and game terms. Ten out of twelve analogues elicited a clear majority of supposed causal-historical responses, even surpassing the proportion of such responses in the corresponding original vignettes featuring proper names, as replicated here, in three cases. The difference between the analogues and the originals did reach statistical significance ($p <$

0.05) in four cases out of twelve.[6] These cases were both analogues of Machery, Mallon, *et al.*, and both analogues of Devitt and Porot's 'Tsu Ch'ung Chih' TVJ-D condition. As for the analogues of Devitt and Porot's 'Tsu Ch'ung Chih' TVJ-D condition, our analogues still elicited about two thirds of the supposed causal-historical responses elicited by the original vignette and probe question featuring a proper name, and even at the lower end of the confidence interval they still elicited close to half of the supposed causal-historical responses elicited by the original vignette and probe question.[7] As for the analogues of Machery, Mallon, *et al.*, whereas the game term analogue only elicited only 31% supposed causal-historical responses, the tool term analogue still elicited about two thirds of the supposed causal-historical responses elicited by the original vignette. Further, although one might argue that the lower end of the confidence interval in both cases need not indicate a problem, the confidence intervals are also compatible with there being two thirds to equal proportions of supposed causal-historical responses, which would certainly be problematic. Moreover, we should keep in mind that the results for individual analogues can be due to accidental features of the vignette in question. For our main argument, it is the overall pattern that counts.

Experiments 1 and 2, taken together, give grounds for serious doubts about the validity of the experimental setup used in existing studies on the reference of proper names. It is a clear shortcoming of such studies that the same setup had never been used on any other class of referring expression, leaving room for uncertainty as to whether the subjects' responses reflected their understanding of proper names, or whether they were driven by some features of the vignettes and probe question that has no bearing on the question of how names refer.

However, it might be objected that we are dismissing the possibility of a causal-historical theory being true of tool and game terms too lightly. We can think of two ways of trying to accommodate our results, by claiming that at least elements of such a theory play a role in determining the extensions of these terms. First, one could hold that, contrary to our assumptions of *prima facie* plausibility, a causal-historical theory *is* true of tool terms and game terms. According to this suggestion, their reference is fixed to whatever shares some fundamental physical traits with 'standard samples' of the relevant tools or games. Arguably,

[6] Using the Bonferroni adjusted alpha level of 0.004 (0.05/12) does not make a difference here either, as the *p*-values in all four cases where a statistical significant difference was found are below 0.004.

[7] Note that, if the causal-historical account indeed is false of the terms used in the analogues, and the analogues elicit two thirds of the supposed causal-historical answers, compared to the originals, it follows that two thirds of all supposed causal-historical answers to the originals cannot count in favour of the causal-historical account. This, in turn, gets the proportion of responses that possibly *can* count in favour of the account to below 50% in all cases.

the best candidates for such fundamental physical traits would be physical structure and/or material composition.[8] Second, one could hold that tool and game terms are initially introduced by a description, but that after the introduction, their reference is passed on in a communicative chain. The term then goes on referring to whatever fits the description with which it was initially introduced, irrespective of which descriptions later speakers associate with the term. Let us call the first view simply 'a causal-historical view' of tool and game terms, and the latter 'a descriptivist reference-borrowing view'.

Both of these views are compatible with a situation where the purported causal-historical referent is the semantic reference, and where all descriptions that a speaker associates with a term are false of the purported causal-historical referent. Hence, were either of them true, that would be a good explanation for the high proportion of supposed causal-historical answers. However, even if either of these views is true, our results would still entail that experimental setups featuring Gödel and Jonah cases fail to measure anything specific to proper names (or natural kind terms, for that matter), as the causal-historical responses might then derive from reference-determining mechanisms that are shared by names and terms for games and tools. Nevertheless, it would still be possible to hold that they measure *something* of relevance to theories of reference, more generally. Experiment 3 was conducted to find out whether such a response is available.

It might additionally be objected that our vignettes are not analogous because proper names are singular terms whereas terms for games and tools are general terms. Notice, however, that when it comes to the standard causal-historical account, the only potentially relevant difference between the two lies in how the terms are introduced in an initial act of baptism. The initial baptismal act has, however, always been left implicit in the vignettes, and this is one of the potential flaws intentionally carried over to our analogues. As such, this difference should not matter. One might additionally claim that, because the initial baptismal act has always been left implicit, these setups only aimed at testing the reference-borrowing part of the causal-historical account, independently of how the term was initially introduced.

Our Experiment 3 goes some way to addressing this objection. It is noteworthy, however, that if one already accepted that all terms can be borrowed, it follows that now these setups can no longer discern between anything of interest. This is so because if all terms can be borrowed, all that remains of interest is how the reference of borrowed terms is fixed.

[8] We do not think such a theory is plausible for terms for tools and even less so for terms for games. Nevertheless, on the assumption that these setups track what they are supposed to track, it *is* a theory that would explain the results of Experiments 1 and 2, and hence one that we wanted to exclude.

Responses to these setups can, however, not discern between different forms of reference fixing because a supposed causal-historical answer is compatible with different forms of reference fixing. If, on the other hand, one did not already accept that all terms can be borrowed, one at least has to admit that there is a plausible alternative explanation for the results, namely that they are due to some feature of the setup that has nothing to do with which theory of reference is true.

Two more things are worth mentioning. First and importantly, we do not take our results to provide positive support for either of our two worries mentioned in the introduction (namely, that subjects in these experimental setups are merely repeating or paraphrasing the narrator of the vignette, or that they base their answers on the reference of the term in their own language). That is, we only take our results to show *that* these setups fail to test what they are supposed to test. Our results do not tell us *why* the setups fail: the two potential problems mentioned would explain why they fail, but there may be other possible explanations. Although we think it is plausible that our results are due to some feature of the setup that has nothing to do with which theory of reference is true, further research will have to show what that feature is. We will not attempt to undertake that here. Hence, at this stage, the worries raised in the introduction merely serve as potential explanations of the results.

Lastly, our results do not touch upon the question of whether there is cross-cultural variation in responses to vignettes modelled on Kripke's Gödel and Jonah case. We only tested responses by Western (specifically: British) participants. However, if we are right, and responses by these participants elicited by setups modelled on Kripke's Gödel and Jonah case do not track what they are supposed to track (that is, whether they assign reference in accordance with the causal-historical account or descriptivism), it follows that any cross-cultural variation found in such setups cannot be variation in whether participants assign reference in accordance with the causal-historical account, but is rather variation in something else.

4. Experiment 3

4.1. Materials and Methods

Experiment 3 was designed to exclude that our results in Experiments 1 and 2 are due to the participants taking either a causal-historical view or a descriptivist reference-borrowing view to be true of terms for games and tools. In order to exclude the first possibility, we presented participants with categorization tasks. In these tasks, we first described the games and tools that featured in the analogues in the same way as we described them in the analogues.

Subsequently, we described a new tool or game that fits the description with which the original tool or game was described, but which is made out of different materials, and/or has a different internal structure. We then asked participants whether the new game or tool was an instance of the game or tool described at the beginning. For the analogues featuring ‘magnometer’, for instance, we used the following task:

> An instrument called ‘magnometer’ is used to measure the strength and direction of magnetic fields. This is the only thing most people know about this instrument. Typically, these instruments work by means of a magnetic core which has copper coils wrapped around it, and they use the properties of electrical current that are affected by the presence of a magnetic field. Recently, it has been discovered that the absorptivity of helium varies with the strength and direction of magnetic fields. Engineers have utilized this effect to create an instrument that measures the strength and direction of magnetic fields by means of measuring helium absorptivity. The instrument they have created works equally well to measure the strength and direction of magnetic fields as devices that use the properties of electrical current.
>
> Accepting that this story is true, please indicate below which one of these two statements you think is true:
>
> The instrument created by the engineers is a magnometer.
>
> The instrument created by the engineers is not a magnometer.

According to the causal-historical view under consideration, the instrument created by the engineers is not a magnometer because it is made of different materials than the standard samples at the time the term was introduced.

It is worth pointing out that in cognitive psychology, categorization tasks have been used to test whether subjects categorize natural kinds according to internal structure such as chemical composition (and thereby in accordance with the causal-historical account), or rather superficial properties such as function (and thereby *not* in accordance with the causal-historical account). However, unlike in the case of setups modelled on Kripke’s Gödel and Jonah cases, similar tasks featuring artefact kinds *have* here been used as a control, and there are systematic differences between the results for artefact terms and natural kind terms (see, for instance, Keil 1989, and more recently Haukioja, Toorman, *et al.* 2023). It is thus plausible to assume that whatever the reason is for getting large proportions of causal-historical responses with our analogues in Experiments 1 and 2, this reason does not carry over to categorization tasks.

We created a categorization task for each analogue, except for the analogues of Li, Liu, *et al.* and the analogue of Devitt and Porot’s Jonah case featuring a tool term. In the case of the former, no categorization task was possible as the relevant descriptions describe an entity as being a winner, and as there can only be one winner, it is not possible to describe another entity as fitting that same description. In the case of the latter, we decided against a categorization task because one of the relevant descriptions concerns a tool being used to identify witches, and we did not want to make assumptions about the reference of terms for non-existing entities. We thus ended up with five different categorization tasks.

The second thing our experiment was designed to do was exclude that purportedly causal-historical responses in Experiments 1 and 2 are the result of participants taking a descriptivist reference-borrowing account to be true of terms for games and tools. As explained above, according to such an account, tool and game terms are introduced by means of a description, after which the term goes on to refer to whatever fits the description with which it was introduced, regardless of which descriptions *current* speakers associate with the term. If participants' responses in Experiments 1 and 2 are due to them taking a descriptivist reference-borrowing account to be true of terms for tools and games, then this would have to be because they understand sentences in the vignettes such as 'Now suppose that these instruments cannot be used to measure the strength and direction of magnetic fields. They are instruments for measuring the spectrum of light.' as stating or implying something about how the term was originally introduced. That is, participants would have to take 'magnometer' to refer to instruments that measure the spectrum of light (the supposed causal-historical referent), regardless of what individual speakers currently believe, because they understand the sentence quoted above as stating or implying that 'magnometer' was initially introduced as a term for instruments that measure the spectrum of light.

To test whether the results in Experiments 1 and 2 are due to the participants taking a descriptivist reference-borrowing account to be true of terms for games and tools, we presented them with the following. First, we repeated the part of the relevant vignette that introduces the hypothetical speakers whose utterances' reference participants are supposed to evaluate (for example, students in astronomy classes in Hong Kong, in case of Devitt and Porot's 'Tsu Ch'ung Chih' vignette, John in Machery, Mallon, *et al.*'s vignette)[9], along with the statements that are to make clear what descriptions they associate with the target term. Subsequently, we transformed statements supposed to indicate that these descriptions are false of the item called by that term into statements about an event in which the term was introduced, and created thereby a clear mismatch between the currently associated descriptions and the description with

[9] Although the questions following Devitt and Porot's 'Tsu Ch'ung Chih' vignette are not explicitly asking about the reference of the target name as used by the hypothetical speakers (i.e., the Hong Kongers), it is still true that the response options Devitt and Porot take to be indicative of the causal-historical account and descriptivism can only be indicative of these accounts if participants themselves are to understand the target name to refer in the same way they take the hypothetical speakers to understand the target name to refer.

which the term was introduced. We then asked the subjects whether the game or tool at issue fits the description with which it was introduced or, rather, the descriptions the hypothetical individual(s) associate with the term. We call these tasks 'descriptivist reference-borrowing tasks'.

For the vignettes used by Devitt and Porot, we used two different versions of this task. In the first version, the hypothetical individuals are included in the same way as they are in the analogues. In the second version, they are not included. In all other tasks, the hypothetical individuals are included in the same way as they are in our analogues. We did this because Devitt and Porot's probe questions do not explicitly ask who the hypothetical speakers are talking about or referring to. As such, if participants' responses are due to them taking a descriptivist reference-borrowing account to be true, this could be because the statements understood as being about an introductory event are understood as being about the beginning of a communicative chain leading to either the hypothetical speakers' use of that term, or to their own communities' use. By varying the presence of the hypothetical speakers we could rule out that participants take a descriptivist reference-borrowing account to be true of the target term as used by members of their own linguistic community as well as by members of the linguistic community of the hypothetical speakers. For the analogue of Devitt and Porot's 'Tsu Ch'ung Chih' TVJ-CH task featuring a term for a game, for instance, participants received the following task:

> Students in agriculture classes in Thailand are told that a board game called 'sheepsy' is a game in which players have to cover the largest portion of the playing field with their stack of sheep. This is the only thing typical Thai people ever hear about this game. However, 20 years earlier 'sheepsy' was introduced as a term for a game in which players have to collect as many sheep as possible.
>
> Accepting that this story is true, please indicate below which one of these two statements you think is true:
>
> Sheepsy is a game in which players have to cover the largest portion of the playing field with their stack of sheep.
>
> Sheepsy is a game in which players have to collect as many sheep as possible.

We created descriptivist reference-borrowing tasks for all analogues except for the analogues of Li, Liu, *et al.* We thus ended up with six different descriptivist reference-borrowing tasks. We did not create a descriptivist reference-borrowing task for the Li, Liu, *et al.* analogues because we took the required explanation to be implausible. In order to explain away the results by means of an appeal to a descriptivist reference-borrowing account, it would have to be assumed that participants take the tool or game at issue to be introduced by means of a description that specifies that the tool or game has won a certain contest. But this is not how tools or games are typically introduced. Moreover, the vignettes name the tool and game prior to any mentions of the outcome of the contest, thereby implying that the names have already been introduced.

One might object that the vignettes in our descriptivist reference-borrowing tasks could be understood as talking about two different tools or games, which happen to have the same name (in this case ‘sheepsy’). As such, the objection goes, they test nothing more than how subjects disambiguate the target term. To be sure, we are not claiming that our tasks are good tests of a descriptivist reference-borrowing account. Rather, our claim is that *if* the original setups test what they are supposed to test, *then* these would be good tests of a descriptivist reference-borrowing account. Our vignettes merely make explicit how subjects would have to understand the vignettes used in our analogues, if their responses are due to them taking a descriptivist reference-borrowing account to be true of terms for tools and games. *If* this makes them into disambiguation tasks, one could indeed argue that answers where participants seemingly take the target term's reference to be determined by the description initially used to introduce the term do *not* indicate that a descriptivist reference-borrowing account is true of the target term. However, similar reasoning could then be used to argue that a supposed causal-historical response to our analogues in Experiments 1 and 2 does not indicate that a descriptivist reference-borrowing account is true of the target terms in the analogues, either. Consequently, if one were to defend the original setups by appealing to the descriptivist reference-borrowing account, then one should expect to find evidence for the descriptivist reference-borrowing account in our tests.

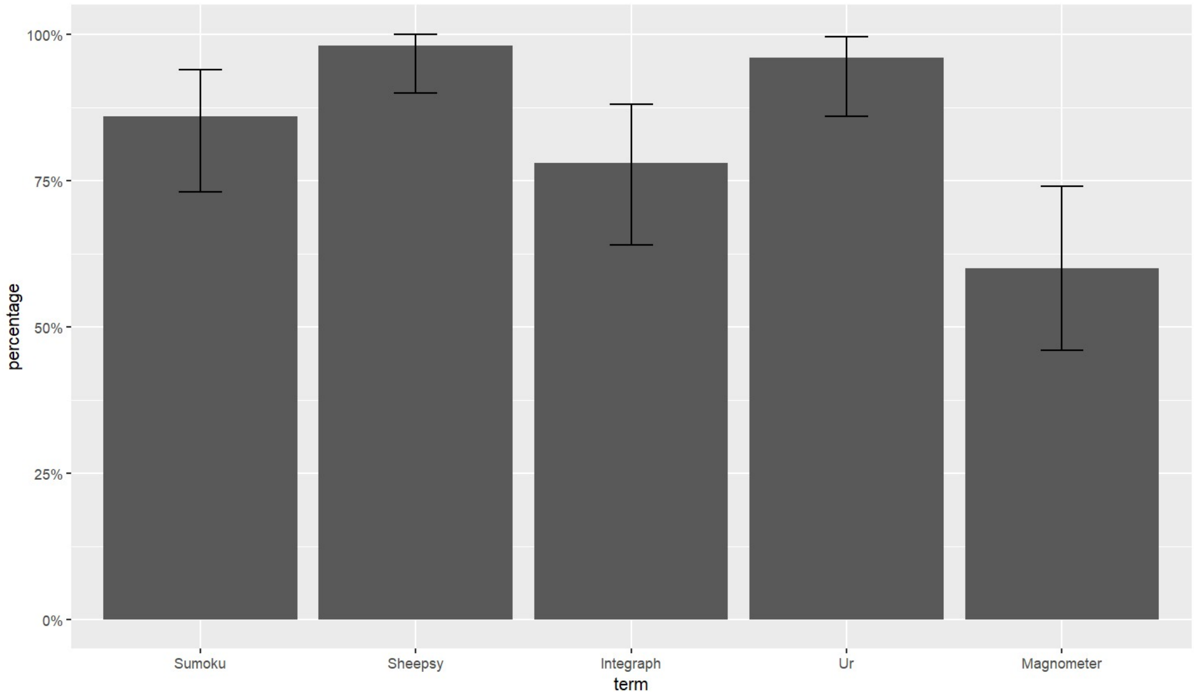
We recruited 103 British participants via the online platform Prolific, all native English speakers. We used the terms ‘integraph’, ‘magnometer’ and ‘twybill’ as terms for tools, from the analogues of Machery, Mallon, *et al.*’s vignette, Devitt and Porot’s ‘Tsu Ch’ung Chih’ vignette and Devitt and Porot’s ‘Ambiorix’ vignette, respectively. We used the terms ‘sumoku’, ‘sheepsy’ and ‘ur’ as terms for games, from the analogues of the same vignettes. After answering mandatory background questions, each participant was randomly assigned to two or three categorization tasks and three descriptivist reference-borrowing tasks. The two types of tasks were switched so that participants never received two tasks of the same type in a row. Additionally, no participant received more than one task featuring the same target term, and all tasks were counterbalanced for order. Moreover, in the categorization task, the response options to the effect that the new item is or is not an instance of the tool or game at issue were switched, so that in half of the cases the first option was that it is an instance of the relevant tool or game, whereas in the other half the first option was that it is not an instance.

4.2. Results

Of the 103 participants, 64 identified as female, 38 as male, and one as neither male nor female. The average age of the participants was 40.00 years (SD: 12.09). In both the categorization task and the descriptivist reference-borrowing task, the percentage of answers incompatible with the causal-historical account or a descriptivist reference-borrowing account, respectively, was compared to chance (50%) by means of a one-sample, two-sided exact binomial test. The results are summarized in Figures 3 and 4, and Tables 3 and 4.

In all cases, participants responded in a way incompatible with the causal-historical account or a descriptivist reference-borrowing account. The difference between the proportion of answers incompatible with the causal-historical account or a descriptivist reference-borrowing account and random responses reached statistical significance in all cases except two, namely the 'magnometer' case in the categorization task and the 'ur' case in the descriptivist reference-borrowing task.

Figure 3: *Percentage of answers to the categorization tasks that are incompatible with the causal-historical account.*

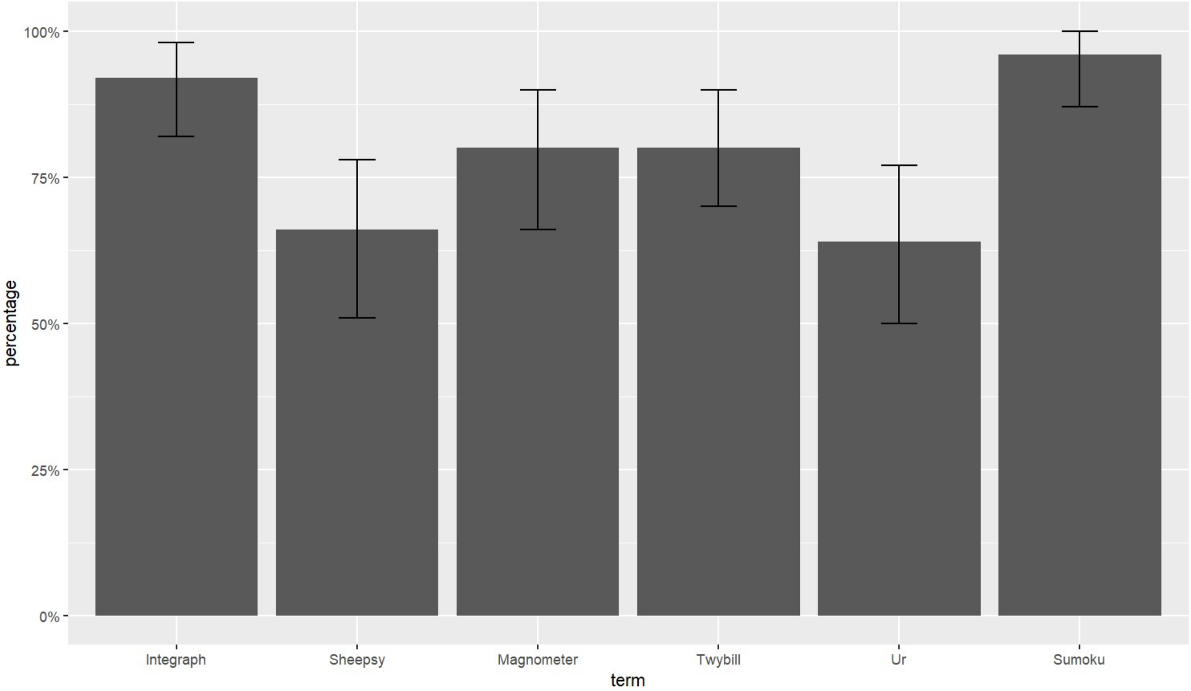


Note: the black lines in the middle of the bar represent the 95% confidence interval of the percentage of answers incompatible with the causal-historical account.

Table 3: *Results categorization task.*

Term:	Percentage 'member' answers:	95% CI:	Independence from chance; p-value:	N:
'Sumoku'	86	73 - 94	<0.001	50
'Sheepsy'	98	90 - 100	<0.001	53
'Integraph'	78	64 - 88	<0.001	50
'Ur'	96	86 - 100	<0.001	50
'Magnometer'	60	46 - 74	0.169	53

Figure 4: *Percentage of answers to the descriptivist reference-borrowing task that are incompatible with a descriptivist reference-borrowing account.*



Note: the black lines in the middle of the bar represent the 95% confidence interval of the percentage of answers incompatible with a descriptivist reference-borrowing account.

Table 4: *Results descriptivist reference-borrowing task.*

Term:	Percentage 'current description' answers:	95% CI:	Independence from chance; p-value:	N:
'Integraph'	92	82 - 98	<0.001	53
'Sheepsy'	66	51 - 79	0.033	50
'Magnometer'	80	66 - 90	<0.001	50
'Twybill'	80	66 - 90	<0.001	50
'Ur'	64	50 - 77	0.053	53
'Sumoku'	96	87 - 100	<0.001	53

4.3. Discussion

Experiment 3 undermines the strategies, mentioned in section 3.3., of responding to Experiments 1 and 2 by assuming that the subjects take the extensions of game and tool terms to be determined in accordance with the causal-historical account, or the descriptivist reference-borrowing account.

One might object to our approach on the grounds that a supposed causal-historical answer merely counts against descriptivism, not in favour of the causal-historical account. Although it might be true that this was Kripke's intention behind his original Gödel case, it is not how experimental philosophers using these setups have understood their work. According to Machery, Mallon, *et al.* (2004: B3-B4), for instance, "The Kripkean intuition is that someone can use the name to speak about the original bearer, whether or not the description is satisfied". Devitt and Porot (2018: 17) take their responses to at least provide "indirect support" for the causal-historical view.

However, given the results of Experiments 1 and 2, and assuming that the causal-historical view is false of terms for tools and terms for games, purported causal-historical answers cannot do this, because such answers are compatible with a theory of reference other than the causal-historical account, namely whatever theory (or theories) is true of terms for games and tools. One might object that the causal-historical account is more plausible for proper names than terms for games and tools. Given what is at stake in this debate, however, this cannot be assumed in advance. Moreover, if we are right and (at least a large proportion of) supposed causal-historical responses are due to some feature of the setup that has nothing to do with which theory of reference is true, one cannot take a supposed causal-historical response to count against descriptivism either.

Lastly, we are not under the illusion that we have excluded all theories of reference that are compatible with a supposed causal-historical answer. One might still make the case that these setups test something of relevance to theories of reference if (1) one can come up with a theory of reference for terms for tools and games that is compatible with a supposed causal-historical response, (2) one can make the case that our analogues are a good test of both this theory as well as of the causal-historical account and (3) this theory of reference is not ad hoc.

However, we take it that the most likely explanation for the large proportion of supposed causal-historical responses we found in Experiments 1 and 2 is that the subjects' responses are based on some feature of the setups that has nothing to do with which theory of reference is true. In any case, even if one can come up with a theory of reference satisfying (1), (2), and (3), one has not yet shown that these setups test what they are supposed to test, but merely that they *may* test something of relevance to theories of reference. Even in that case, however, one at least has to admit that our preferred explanation provides an alternative that has not yet been ruled out. We take the burden of proof to be on the defender of these setups.

5. Conclusion

Our experiments strongly suggest that the widely used experimental setups modelled on Kripke's Gödel and Jonah cases do not tell us much about whether or not test subjects assign reference in accordance with the causal-historical account. Experiments 1 and 2 show that subjects give relatively large proportions of supposed causal-historical responses in setups that are otherwise closely analogous, but feature terms for which the causal-historical account is arguably not true. Experiment 3 reinforces the conclusion by ruling out two attempts at explaining away the results of Experiments 1 and 2.

Acknowledgements

We are very grateful to Lucien Baumgartner, Hedwich Boelens, Daniel Cohnitz, Mons Nyquist, Wojciech Rostworowski, and two anonymous referees for comments on earlier versions of this paper. We are also grateful to the audience at the Third European Experimental Philosophy Conference in Zurich in 2023 and to the participants in an online LOGOS reading group on experimental philosophy. Special thanks go to Genoveva Martí for inviting us to present this work at the latter, and to Michael Devitt for lengthy email discussions.

ORCID

Jeske Toorman 0000-0001-5700-8335

Jussi Haukioja 0000-0002-3906-8278

References

Deutsch, Max (2009) 'Experimental Philosophy and the Theory of Reference', *Mind and Language* **24**: 445–66. doi: 10.1111/j.1468-0017.2009.01370.x

Devitt, Michael and Nicolas Porot (2018) 'The Reference of Proper Names: Testing Usage and Intuitions', *Cognitive Science* **42**: 1552–85. doi: 10.1111/cogs.12609

Domaneschi, Filippo, Massimiliano Vignolo, and Simona Di Paola (2017) 'Testing the Causal Theory of Reference', *Cognition* **161**: 1–9. doi: 10.1016/j.cognition.2016.12.014

Domaneschi, Filippo and Massimiliano Vignolo (2020) 'Reference and the Ambiguity of Truth-Value Judgments', *Mind & Language* **35**: 440–55. doi: 10.1111/mila.12254

Haukioja, Jussi, Jeske Toorman, Giosuè Baggio, and Jussi Jylkkä (2023) ‘Are Natural Kind Terms Ambiguous?’, *Cognitive Science* **47:** e13335. doi: 10.1111/cogs.13335

Keil, Frank C (1989) *Concepts, Kinds and Cognitive Development*. MIT Press.

Kripke, Saul A (1980) *Naming and Necessity*. Basil Blackwell.

Li, Jincai, Longgen Liu, Elizabeth Chalmers, and Jesse Snedeker (2018) ‘What Is in a Name?’, *Cognition* **171**: 108–11. doi: 10.1016/j.cognition.2017.10.022

Ludwig, Kirk (2007) ‘The Epistemology of Thought Experiments’, *Midwest Studies in Philosophy* **31**: 128–59. doi: 10.1111/j.1475-4975.2007.00160.x

Machery, Edouard (2011) ‘Variation in Intuitions about Reference and Ontological Disagreements’, in Steven D Hales, ed., *A Companion to Relativism*: 118–36. Wiley-Blackwell. doi: 10.1002/9781444392494.ch7

Machery, Edouard (2024) ‘Experimental Philosophy of Language: Proper Names and Predicates’, in Alexander M Bauer and Stephan Kornmesser, eds., *The Compact Compendium of Experimental Philosophy*: 183–210). de Gruyter. doi: 10.1515/9783110716931-009

Machery, Edouard, Ron Mallon, Shaun Nichols, and Stephen P Stich (2004), ‘Semantics, Cross-Cultural Style’, *Cognition* **92**: B1–B12. doi: 10.1016/j.cognition.2003.10.003

Machery, Edouard, Christopher Y Olivola, and Molly de Blanc (2009), ‘Linguistic and Metalinguistic Intuitions in the Philosophy of Language’, *Analysis* **69**: 689–94. doi: 10.1093/analys/anp095

Machery, Edouard, Justin Sytsma, and Max Deutsch (2015) ‘Speaker’s Reference and Cross-Cultural Semantics’, in Andrea Bianchi, ed., *On Reference*: 62-76. Oxford University Press.

Machery, Edouard, Christopher Y Olivola, Hyundeuk Cheon, Irma T Kurniawan, Carlos Mauro, Noel Struchiner, and Harry Susianto (2023), ‘Is Identity Essentialism a Fundamental Feature of Human Cognition?’, *Cognitive Science* **47**: e13292. doi: 10.1111/cogs.13292

Mallon, Ron, Edouard Machery, Shaun Nichols, and Stephen Stich (2009) ‘Against Arguments From Reference’, *Philosophy and Phenomenological Research* **79**: 332–56. doi: 10.1111/j.1933-1592.2009.00281.x

Martí, Genoveva (2009) ‘Against Semantic Multi-Culturalism’, *Analysis* **69**: 42–48. doi: 10.1093/analys/ann007

Sytsma, Justin and Jonathan Livengood (2011) ‘A New Perspective Concerning Experiments on Semantic Intuitions’, *Australasian Journal of Philosophy* **89**: 315–32. doi: 10.1080/00048401003639832

Sytsma, Justin, Jonathan Livengood, Ryoji Sato, and Mineki Oguchi (2015) ‘Reference in the Land of the Rising Sun: A Cross-cultural Study on the Reference of Proper Names’, *Review of Philosophy and Psychology* **6**: 213–30. doi: 10.1007/s13164-014-0206-3

van Dongen, Noah, Matteo Colombo, Felipe Romero, and Jan Sprenger (2021) ‘Intuitions About the Reference of Proper Names: a Meta-Analysis’, *Review of Philosophy and Psychology* **12**: 745-74. doi: 10.1007/s13164-020-00503-8