Slurs in quarantine

Bianca Cepollaro (San Raffaele), Simone Sulpizio (Milan-Bicocca), Claudia Bianchi (San Raffaele) and Isidora Stojanovic (Institut Jean Nicod)

Preprint version of the article forthcoming in *Mind & Language* (2023)

Abstract

We investigate experimentally whether the perceived offensiveness of slurs survives when they are reported, by comparing Italian slurs and insults in *base utterances* (Y is an S), *direct speech* (X said: “Y is an S”), *mixed quotation* (X said that Y is “an S”), and *indirect speech* (X said that Y is an S). For all strategies, reporting decreases the perceived offensiveness without removing it. For slurs, but not insults, *indirect speech* is perceived as more offensive than *direct speech*. Our hypothesis is that, because slurs constitute hate speech, speakers employ quotation marks to signal their dissociation from slur use.

KEYWORDS

expressives, slurs, insults, hate speech, quotation, speech reports

1. INTRODUCTION

When philosophers and linguists turned their attention to slurs—derogatory terms that target people on the basis of their belonging to certain groups, determined by nationality, ethnicity, gender, sexual orientation, and the like[1]—one question that they raised was whether the perceived offensiveness of slurs survives under quotation. This issue—let us call it the *survival question* —is interesting for both theoretical and practical reasons. From a theoretical point of view, studying the offensive

[1] See, among others, Hornsby (2001), Potts (2007), Hom (2008), Hom and May (2013), Bolinger Jorgensen (2017), McCready and Davis (2017), Jeshion (2018), Nunberg (2018), Camp (2018), Stojnić and Lepore (2022). See Hess (2022) for a detailed survey of the debate.

potential of slurs across contexts offers some insight into their nature, by providing valuable clues as to how their derogatory power should be analyzed. From a practical perspective, it is important to establish whether and how slurs can be mentioned. This is an ethical issue, but such a normative determination should be based on the observation of whether and how the behavior of slurs changes from one use to another.

In this paper, we take seriously the idea that the vexed question of how slurs behave under quotation is, to a certain extent, empirical and we thus assess it experimentally. We focus on the phenomenon of *reported* slurs—which, as we will see, encompasses a range of strategies. Our goal is to provide empirical results that could feed the work of philosophers of language and linguists, on the theoretical side, and ethicists and policymakers, on the practical side. In this section, we start by giving a brief survey of the debate on reported slurs, then formulate three questions on reported slurs that we want to assess empirically. In Sections 2 and 3 we present our two studies. We end by discussing how our results impinge on the current debate.

In the philosophical literature on slurs there is no consensus on how speech reports affect the perceived offensiveness of slurs. First, scholars do not agree on *whether* the offensiveness survives in reports, or on how to diagnose whether it does, or on what its alleged survival would mean for a theory of slurs. On the one hand, quoted slurs are often seen as fairly inoffensive. This should not be surprising. After all, quoted terms should be inert. When a speaker uses a slur, their speech act is (normally) derogatory because they are referring to the slurred group in a derogatory way.[2] But when the slur is only reported, the derogatory character of the slurring act should, in principle, be only ascribed to the reported speaker, and not to the reporter. Mentioning a term is not the same thing as using it and reported slurs should not pose the same problems as ones that are being used. Several authors seem to agree with this intuition (Hornsby, 2001, pp. 129-130; Potts, 2005, pp. 161-62, Hom, 2008, p. 16; Williamson, 2009, pp. 142-143). Bolinger Jorgensen writes:

> The least controversial inoffensive occurrences of slurring terms are cases where the terms are only mentioned. In direct quotation, or when some contextual constraint (e.g. a hearer's insistence to 'tell me exactly what he said') leaves the speaker with no alternative to

[2] We add the qualification "(normally)" because we will set aside reclaimed uses, as well as any other possible non-derogatory uses.

> mentioning an offensive term, the offensiveness of the term remains embedded in its original context. Our indignation, if it is aroused, is directed at the individual whose utterance is being quoted, rather than the current speaker. Similarly, when a slurring term is mentioned in a dictionary entry (Bolinger Jorgensen, 2017, p. 442).

Hess concurs:

> (4) “Wop” is a slur for Italians.
>
> (5) My child said our neighbor was a “jap”. I had to do a lot of explaining.
>
> … In (4) and (5) the expressions are quoted, and therefore only *mentioned,* and not actually *used.* Quotation neutralizes derogation and offensiveness, making it possible to talk *about* slurs in a neutral way. … The mechanisms of quotation and meta-linguistic negation are external to a theory of slurs—whatever the nature of their derogatory meaning, it is to be expected that it becomes neutralized in such contexts” (Hess, 2020, p. 89).

On the other hand, some scholars maintain that the perceived offensiveness of slurs has little to do with whether they are mentioned or used (see Anderson & Lepore, 2013a, 2013b; Anderson, 2016). And some acknowledge that at least *something* in slurs’ offensiveness survives even when they are quoted. Indeed, Bolinger Jorgensen hedges her previous claim as follows:

> Perhaps you think, as I do, that there is still something strange (or offensive) about listing each of the slurs explicitly rather than giving a blanket admonition to avoid slurring terms. … [T]he offensiveness of slurs projects out of various forms of embedding, including indirect reports, negations, and mentions. (Bolinger Jorgensen, 2017, pp. 439-452)

In this work, we assess the *survival question* on empirical grounds, and articulate it into two further queries: the *gradability question* and the *strategy question*.

The *gradability question* asks whether the perceived offensiveness of slurs is gradable; that is, whether certain environments (as, in the present case, speech reports) can affect the perceived offensiveness of slurs in a way that admits intermediate states between no effect (that is , reported slurs are equally offensive as non-reported slurs) and full cancellation (reported slurs are not offensive at all). There is no widespread agreement in the literature on the *gradability question*, but many scholars suggest that offensiveness *is* gradable. Bolinger Jorgensen suggests that when a given slur is mentioned rather than used, its offense potential is *diminished*:

> While uses are associated with hostile, aggressive and threatening behavior, contextually inappropriate mentions appear to be associated with *tamer* (though not benign) attitudes, ranging from simple insensitivity to perverse pleasure at saying discomfiting words, and disregard for the risk of encouraging derogating uses of the slur. … [In certain contexts where slurs are mentioned] offensiveness is far *less severe*, *but not wholly absent* … (contra the prediction of a mere mention/use distinction) … [T]he mere fact that the slurs are mentioned rather than used does not fully neutralize the offense potential. (Bolinger Jorgensen, 2017, p. 452; our italics; see also Davis & McCready, 2020, p. 73 for a similar point.)

Others, however, believe that the offensiveness of slurs is *not* gradable: "[Slurs'] power to offend does not depend on malicious intent on the part of individual speakers, and so can be not at all mollified by scholarly appeals to use/mention distinctions" (Sullivan, 2022, p. 1479).[3]

Finally, the *strategy question* asks whether different strategies of reporting other people's slurring utterances affect the perceived offensiveness of the resulting utterance in different ways. Some scholars seem to suggest that this is the case. In the passage cited above, for instance, Bolinger Jorgensen (2017, pp. 442, 444) writes that "[i]n *direct* [speech] …, the offensiveness of the term remains embedded in its original context", and that "the offensiveness of slurs is not insulated by embedding in *indirect* [speech]s" (our emphasis). In our studies, for the first time, the *strategy question* is explicitly

[3] Stojnić and Lepore (2022, pp. 2, 7) suggest that the emphasis on the gradability of the offensiveness of slurs was improperly used to argue that the feeling that the offensiveness survives in reports is merely residual: "The literature has focused on uses of slurs with an emphasis on variation in offensiveness—e.g., derogatory uses are deemed more offensive than pedagogical ones. Such considerations tend to emphasize the role of reasoning about typically present, or conventionally signaled, speaker intentions, and semantic/pragmatic mechanisms which, in various conversational contexts or linguistic environments, block the expression of, and the speaker's commitment to, the postulated pejorative content, which is often taken to express, or otherwise commit the speaker to, a negative attitude. As a result, there's a tendency to treat the offensiveness of slurs in meaning-attributive and quotative environments as "derivative" or "residual," and of less severity than "direct" uses. It is not implausible that there is gradation in severity of transgression among various acts of slur-tokenings. But any approach that treats uses intended to denigrate as explanatorily central, while corresponding displays as marginal, is mistaken. … According to these explanations, mentioned slurs are invariably less offensive than (nonreclaimed) used ones, and pejorative effects that mentionings carry are derivative from their uses. This is a mistake".

formulated and empirically assessed . We have various non-equivalent strategies at our disposal to report the words of others. Given (A), an utterance like “Y is an S” (S is a slur and Y a person’s name), uttered by Z, we will consider three ways of reporting it (B, C, D):

(A) *Base utterance*: Z: Y is an S.
(B) *Direct speech*: Z: X said: “Y is an S”.
(C) *Mixed quotation*:[4] Z: X said that Y is “an S”.
(D) *Indirect speech*: Z: X said that Y is an S.

Each reporting strategy is characterized by distinctive features; moreover, not all are available for both written and spoken language.[5] Oral language, for instance, does not have quotation marks at its disposal.[6] In what follows, we focus on written language only, and consider the aforementioned three ways to report words—(B) *Direct speech*, (C) *Mixed quotation*, and (D) *Indirect speech*—such that (B) and (C) include quotation marks (“”), while (D) employs constructions with a verbum dicendi (said that) with no quotation marks. Our goal is to explore whether the perceived offensiveness of slurs is affected by each kind of strategy and how—something that has not yet been done in the literature.

We believe that these three questions (*survival*, *gradability*, and *strategy*) cannot be entirely answered from the philosopher’s armchair, for they are, at least to some extent, empirical. They amount to asking how slurs are *perceived* when they occur in reported discourse. Hence, theoretical investigation should take into account empirical data.

At this point, some clarifications are in order. Some authors distinguish between derogation and offense. In the philosophical jargon, derogation roughly has to do with what slurs conventionally do,

[4] We do not need to distinguish here between *mixed* and *scare quotation*, for both would work as “distancing” devices, as Belleri (2020, p. 22) puts it.

[5] See, among others, Davidson (1979), Predelli (2003), Saka (2003), Maier (2014), Cappelen, Lepore, and McKeever (2020).

[6] Occasionally, oral language resorts to alternative means to explicitly signal quotation, for instance oral promptings like “quote-unquote”, or the finger quotes, or a special intonation, but neither is necessary. Nor are quotation marks systematically needed in written language (think for instance of italicization). As it has been noted (Cappelen, Lepore & McKeever, 2020), when a speaker says (orally or in writing), “My name is Julia”, they do not need (if talking) to say “My name is—quote-unquote—Julia”, nor to use a special intonation, nor to add finger quotes; and (if writing), they do not need to have “My name is ‘Julia’”.

for example, what they convey about their targets. Offense, in contrast, is the psychological reaction of discomfort that some people have when exposed to a slur. Most philosophers agree with the claim that slurs reported in quotation marks are not *derogatory*, for they cannot convey anything at all about the target group, since the term is not *used*; however, they can still trigger *offense.* Thus, Jeshion (2020, p. 110) writes: "[A reported slur] exemplifies a speaker-non-derogating use. ... While [reported slurs] … can be used non-derogatorily, … the very utterance [of certain slurs] by a non-member of the slurred group is regularly perceived to be an offense". Similarly, David and McCready (2020, p. 69) note: "[Slurs can be] non-derogatory and nevertheless offensive, a subtle distinction that we think is useful in understanding some of the debate and misunderstanding that arises in discussions surrounding the (in)appropriate use and/or mention of slurs".

What offense is, and by what it is caused, varies very much from author to author. But is it possible to know whether the intuitions of ordinary speakers regarding the perceived offensiveness of slurs track derogation or, rather, offense? Our study remains neutral with respect to the issue of what is the offensive effect that reported slurs have: even if our experimental question features the notion of "offensive", our design does not distinguish between the various layers identified in the philosophical debate (viz. derogation, offense, and possibly more). We opted for using "offensive" for two complementary reasons. First, it is a very commonly used concept with which Italian speakers are familiar: "*offensive*" is by far the most frequent and used notion compared to alternatives such as "*denigratorio*" ("derogatory") (Google lists seven million occurrences for the former vs. 120,000 occurrences for the latter). The second reason is that, unlike most alternatives, offensiveness has been used in other rating studies on taboo words (including slurs and insults, e.g., Eilola & Havelka, 2010; Janschewitz, 2008; Rosenberg, Sikström & Garcia, 2017; Sulpizio et al., 2019). This said, it is reasonable to expect that participants interpreted "offensive" in the ordinary sense rather than as philosophical jargon which would be contrasted with "derogatory". What this study provides is thus an indispensable contribution to tackle the disagreement concerning the data: given an intuitive and ordinary understanding of "being offensive", do slurs that are being merely reported display this feature? Testing participants and gathering data about this is the first step to address the vexed question of reported slurs.[7]

[7] We thank an anonymous reviewer for urging us to clarify this point.

An additional reason to test perceived offensiveness is that this allowed us to directly compare our results to those reported by Cepollaro, Bianchi and Sulpizio (2019) —which is, to our knowledge, the only existing experiment on reporting slurs—by asking participants the very same question.[8] In that study, the authors compared *base utterances* (Z: Y is an S) with *indirect speech* (Z: X said that Y is an S), for two categories of pejoratives: slurs (derogatory terms that targets people for their belonging to certain social groups) and so-called particularistic insults (Saka 2007) —or just "insults"—that is, pejorative terms that target individuals without referring to recognized social groups (such as "asshole" or "jerk"). They found that *indirect speech* (Z: X said that Y is an S) can decrease—without deleting—the offensiveness of *base utterances* (Z: Y is an S) to a similar extent when S is a slur or an insult. Their findings provide some preliminary answers to the questions in which we are interested. In particular, for the *survival question*, Cepollaro, Sulpizio and Bianchi (2019) show that at least *indirect speech* is able to somewhat decrease the offensive force of utterances that contain slurs and insults—without deleting it entirely—, but they provide no data for other discourse reporting strategies, especially those involving quotation marks. As for the *gradability question*, their study suggests that the perceived offensiveness of both slurs and insults is gradable in the sense that it admits intermediate degrees between when these terms are used and when they are reported. Finally, Cepollaro, Sulpizio and Bianchi (2019)'s results do not provide any data about the *strategy question*, for they only considered *one* way of reporting other people's words, namely *indirect speech* under a verbum dicendi.

In this work, we go beyond Cepollaro, Sulpizio and Bianchi (2019)'s experiment by introducing two new conditions which—unlike *indirect speech* (D) —involve quotation marks: (B) *Direct speech* and (C) *Mixed quotation*. In Study 1, we compare *direct* and *indirect speech* (B and D). In Study 2, we compare *direct speech* and *mixed quotation* (B and C). We test whether and to what extent each of these discourse reporting strategies affects the perceived offensiveness of slurs and insults. In doing so, not only do we address the *survival question* (does the offensiveness of slurs survive when they are reported rather than used?) and the *gradability question* (does the offensiveness of slurs admit intermediate degrees?) for all these reporting tools, but we also assess, for the first time, the *strategy*

[8] Panzeri and Carrus (2016) also measured the perceived offensiveness of reported slurs and found that indirect speech decreases it without deleting it, but they lacked any condition with base utterances. Instead, they directly compared reported slurring utterances of the form "Y said that X is an S" with unembedded slurs like "S". One interesting aspect of their study, though, is that they considered various environments, including negation, antecedents of conditionals, indirect speech, and questions.

question (do different discourse reporting strategies affect the perceived offensiveness of slurs in different ways?). In both experiments, we look not only at slurs, but also at particularistic insults, as they can help us disentangle what depends specifically on the distinctive features of slurs from what concerns pejoratives in general. Consequently, our findings are relevant not only to the debate on slurs, but also to the broader literature on expressive meaning.

2. STUDY 1

2.1 Methods

2.1.1 Participants

Participants were recruited via university student mailing lists or social networks and provided with a link to the online survey (that were also invited to send the link to their friends and acquaintances). Participants were asked not to take part to the study if they had taken part in a similar study in the past. Eighty-four volunteers took part in the study. Fourteen participants were excluded from the analyses as they were not native speakers of Italian. The final sample consisted of 70 participants (43 females; M_{age} = 22.41, SD = 5.67), all Italian native speakers.

2.1.2 Materials and procedure

We used the same stimuli as Cepollaro, Sulpizio and Bianchi (2019) that included 13 stimuli for slurs (e.g., *crucco*, kraut), 13 for the corresponding non-slurring category labels (e.g., *tedesco*, German) and 13 for particularistic insults (e.g., *idiota*, idiot). These stimuli were such that: a) for each item in the slur list, there was a corresponding item in the non-slurring label list. For each social category, there was only one pair <slur, corresponding non-slurring label>: for instance, despite the fact that there are many different slurs in Italian for prostitutes, only one slur and one corresponding non-slurring counterpart were selected;[9] b) slurs and insults were matched on offensiveness (M_{slurs} = 4.70 vs. $M_{insults}$ = 4.63, $t < 1$, $p > .6$; frequency of written occurrence ($p > .06$; frequency was extracted by SUBTLEX-IT, a word frequency database based on subtitles of Italian movies and tv series, Crepaldi, Keuleers, Mandera & Brysbaert, 2013) and length in letters ($p > .9$) —which are known to significantly impact on

[9] There were two exceptions for gay men and people with disabilities. In these cases, Cepollaro, Sulpizio and Bianchi's (2019) stimuli included two pairs <slur, non-slurring label> for the same group: <"frocio", "omosessuale"> and <"finocchio", "gay"> for gay men; <"handicappato", "disabile">, <"mongoloide", "down"> for people with disabilities.

recognition and reading processes (e.g., Kliegl, Grabner, Rolfs & Engbert, 2004).

Stimuli were embedded in three types of sentences: (A) *Base utterances* of the form *Z: X is a P* (e.g., *Umberto*: *Vittoria è crucca*, Umberto: Vittoria is a kraut); (B) *Direct speech*, of the form *Z: Y said: "X is a P"* (e.g., *Benedetta: Umberto ha detto: "Vittoria è crucca"*, Benedetta: Umberto said: "Vittoria is a kraut"); and (D) *Indirect speech* of the form *Z: Y said that X is a P* (e.g., *Benedetta: Umberto ha detto che Vittoria è crucca,* Benedetta: Umberto said that Vittoria is a kraut). In all, X, Y and Z are proper names (roughly half male, half female), and P is a predicate: either a slur, or an insult, or a non-slurring category label. Moreover, to introduce variability into the content and structure of the stimuli, 15 filler sentences were included; these sentences contained no offensive term and had a different syntactic structure (e.g., utterances like "*Stella ha una custodia da pc molto costosa*", "Stella has a very expensive computer case").

Sentences were presented in a random order and all participants were presented with all stimuli.

Participants were asked to rank the offensiveness of the reporter's utterance on a 7-point scale from 1 (not at all offensive) to 7 (highly offensive). For instance, after the stimulus "*Umberto*: *Vittoria è crucca*", (Umberto: Vittoria is a kraut), they were asked "*Quanto è offensiva la frase pronunciata da Umberto?*" ("How offensive is the sentenced uttered by Umberto?"). That is, each item was followed by a question specifically referring to the utterance of that item's speaker. Finally, participants provided demographic information (age, gender, native language). The questionnaire was created and administered with SoSci Survey (Leiner, 2019).

2.2 Results

For both Study 1 and 2, the statistical analyses were conducted in R (R Core Team, 2015), using the ez library (version 4.4-0, Lawrence, 2016). Figure 1 shows the average offensiveness rates for the different types of words. To evaluate the impact of different types of embedding sentences on the three types of words, a 3 (Type of Word: slurs vs non-slurring labels vs insults) x 3 (Type of Sentence: *base utterance* vs *direct speech* vs *indirect speech*—A vs B vs D) ANOVA was run. Both factors were within-participants. Where appropriate, critical values were adjusted using the Geisser and Greenhouse (1959) correction for violation of the assumption of sphericity.

The results showed main effects of Type of word (F (2, 138) = 175.15, p <.001) and Type of Sentence (F (2, 138) = 41.26, p <.001). More interestingly, the interaction was significant (F (4, 276) =

25.89, p <.001). Follow-up multiple comparisons (with Bonferroni correction) showed that both non-slurring labels[10] and insults were more offensive in (A) *Base utterances* than in (B) *Direct speech* (labels: p =.001; insults = p <.001) and (D) *Indirect speech* (labels: p <.001; insults: p <.001), but no difference emerged between *direct* and *indirect speech* (both ps >.9). By contrast, for slurs, while (A) *Base utterances* were again more offensive than both (B) *Direct* and (D) *Indirect speech* (both ps < . 001), (D) *Indirect speech* was more offensive than (B) *Direct speech* (p = .01).

Finally, an ancillary analysis directly compared filler sentences with sentences containing non-offensive labels (in the *base utterance* condition) by means of a paired t-test, showing that the latter received higher offensiveness rates than the former (mean offensiveness for fillers = 1.12 and for non-offensive labels = 2.16; t (69) = 11.32, p < .001).

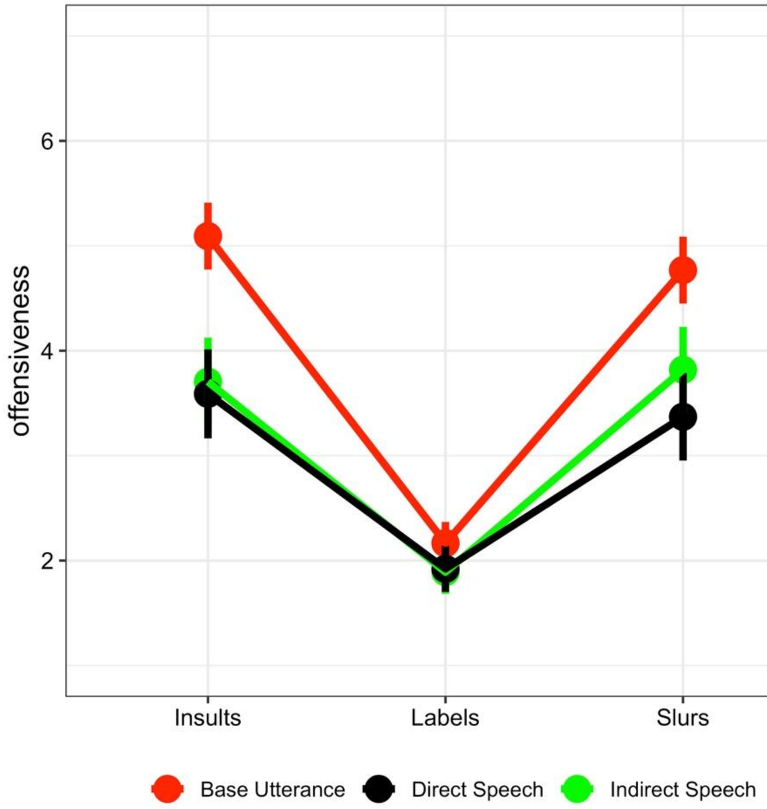

Figure 1. Mean values of offensiveness by condition for Study 1. Points represent participants mean scores. Error bars indicate 95% confidence intervals across individuals.

[10] Even if the rates of perceived offensiveness were very low for non-slurring category labels, as the figure clearly shows.

2.3 Discussion

It is not surprising that discourse reports (both *direct* and *indirect speech*) featuring slurs and insults are perceived as *less* offensive than base utterances featuring derogatory language: The speaker reports someone else's offensive words, rather than introducing them themselves, and this leaves open the possibility that someone *other* than the speaker is responsible for using offensive language.

What is more interesting is that, for slurs, the strategy employed to quote someone else's words *matters* to how offensive the resulting utterance is perceived. As we have seen, *indirect speech* (Z: X said that Y is an S) was rated more offensive than *direct speech* (Z: X said: "Y is an S"). There are two questions that arise here: why does *indirect speech* get higher rates of perceived offensiveness than *direct speech*? And why does this happen for slurs, but not for non-slurring insults? Let us start from the latter question. Our hypothesis is that slurs (terms that derogate social groups) expose the speaker who utters them to higher social risks than non-slurring insults (terms that derogate individuals without any reference to social groups). The use of slurs is often socially if not legally sanctioned in a way that non-slurring insults are not, for slurs are usually seen as paradigmatic instances of hate speech, that is, discourse that erodes social equality and the social fabric (see, among others, Delgado, 1982; Matsuda, 1989; Lawrence III, 1992; Butler, 1997; Waldron, 2010; Brown, 2017; Tirrell, 2021). Non-slurring terms, in contrast, are usually taken to express the speaker's feelings or negative judgement about the target, regardless of their social identity.[11] We hypothesize that speakers are thus especially cautious about using slurs, even when they are reported: The risk they want to avoid, we suppose, is passing as bigots. Resorting to quotation marks is a way to explicitly signal that the speaker is reporting someone else's words. However, even showing that one is willing to take the risk to pass as a bigot suffices sometimes to make one look like a bigot. As Camp puts it:

> In the case of slurs in particular, given how loaded and repugnant they are, a reporting speaker's failure to strongly distance themselves from the perspective often constitutes positive evidence of endorsement. Given this, a reporting speaker may rightly be held responsible for not

[11] Slurring terms are direct indicators of hate speech, but of course, hate speech can also be performed by means of non-slurring insults rather than slurs, as in "Italians are assholes" and the like. See the theoretical and empirical literature on the harmful uses of social generics (among others, Gelman, Ware & Kleinberg, 2010; Rhodes, Leslie & Tworek, 2012; Wodak, Leslie & Rhodes, 2015; Leslie, 2017; Saul, 2017; Ritchie, 2019).

> having explicitly repudiated it or engaging in circumlocution—*even* for direct quotation" (Camp, 2018, p. 52).[12]

But let us go back to our first question, namely why the risk of coming across as a bigot would be higher with *indirect* rather than *direct speech*. That is straightforward. Take (1) and (2):

(1)Roberta: Martina said that Giovanni is a fag.
(2)Roberta: Martina said: "Giovanni is a fag".

In the absence of contextual information, we do not know whether in (1) Martina slurred Giovanni and Roberta is merely reporting it, or whether Martina only said Giovanni was gay and Roberta reported her words with a slurring utterance. Thus (1) is ambiguous: On one reading, Roberta comes across as a homophobe who introduces herself the slur; on another, she only reports a slur used by the reportee. When, in contrast, a speaker chooses to employ quotation marks as in (2), they are *explicitly* signaling that the following words are not their own, but someone else's.

When a socially risky term like a slur is used in a report, speakers must resort to a non-ambiguous quotational device such as written marks to manifest their *verbatim* report of the offensive term. In the case of insults, in contrast, this does not seem to be necessary: *Direct* and *indirect speech* decrease the perceived offensiveness of insults to a similar extent.

With these results in hand, we now investigate whether different discourse reporting strategies that employ quotation marks affect the offensiveness of slurs and insults in different ways. More precisely, we want to know whether the variation in the perceived offensiveness of slurs is sensitive to where exactly quotation marks are put, whether on the whole quoted utterance (like in *direct speech*, Z said that "Y is an S"), or whether on the slur only (like in *mixed quotation*, Z said that Y is "an S").

[12] See also Bolinger Jorgensen (2017, p. 452); compare to Harris and Potts (2009, pp. 546-547), Lasersohn (2007, p. 228).

3. STUDY 2

3.1 Method

3.1.1 Participants

Participants were recruited via university student mailing lists or social networks and provided with a link to the online survey (that they were also invited to send to their friends and acquaintances). As in Study 1, participants were asked not to take part in the study if they had taken part in a similar study in the past, so that each participant could not take part in both studies. Seventy-five volunteers took part in the online study. Seven participants were excluded from the analyses as they did not provide demographic information. The final sample consisted of 68 participants (43 females; M_{age} = 25.451, SD = 10.35), all Italian native speakers.

3.1.2 Materials and procedure

The same words of Study 1 were used and embedded in three types of sentences: (A) *Base utterances* of the form *Z: X is a P* (e.g., *Umberto*: *Vittoria è crucca*, Umberto: Vittoria is a kraut); (B) *Direct speech*, of the form *Z: Y said: "X is a P"* (e.g., *Benedetta: Umberto ha detto: "Vittoria è crucca"*, Benedetta: Umberto said: "Vittoria is a kraut") (like Study 1); and (C) *Indirect speech* of the form *Z: Y said that X is "a P"* (e.g., *Benedetta: Umberto ha detto che Vittoria è "crucca"*, Benedetta: Umberto said that Vittoria is "a kraut").

The procedure was identical to that of Study 1.

3.2 Results

Figure 2 shows the average offensiveness rates for the different types of words. To evaluate the impact of different types of embedding sentences on the three types of words, a 3 (Type of Word: slurs vs non-slurring labels vs insults) x 3 (Type of Sentence: *base utterance* vs *direct speech* vs *mixed quotation*— A vs B vs C) ANOVA was run. Both factors were within-participants. Where appropriate, critical values were adjusted using the Greenhouse and Geisser (1959) correction for violation of the assumption of sphericity.

The results showed main effects of Type of word (F (2, 134) = 169.84, p <.001) and Type of Sentence (F (2, 134) = 69.14, p <.001). More interestingly, the interaction was significant (F (4, 268) =

60.80, p <.001). Follow-up multiple comparisons (with Bonferroni correction) showed the same pattern for all three types of words: They were all perceived as more offensive in (A) *Base utterance* than in (B) *Direct speech* (labels: p =.01; insults: p <.001; slurs: p < .001) and (C) *Mixed quotation* (labels: p =. 002; insults: p <.001; slurs: p <.011), but no difference emerged between (B) *Direct speech* and (C) *Mixed quotation* (both ps >.9).

Finally, just like we did for Study 1, an ancillary analysis directly compared filler sentences with sentences containing non-offensive labels (in the *base utterance* condition) by means of a paired t-test, showing once again that the latter received higher offensiveness rates than the former (mean offensiveness for fillers= 1.08 and for non-offensive labels = 2.08; t (67) = 10.41, p < .001).

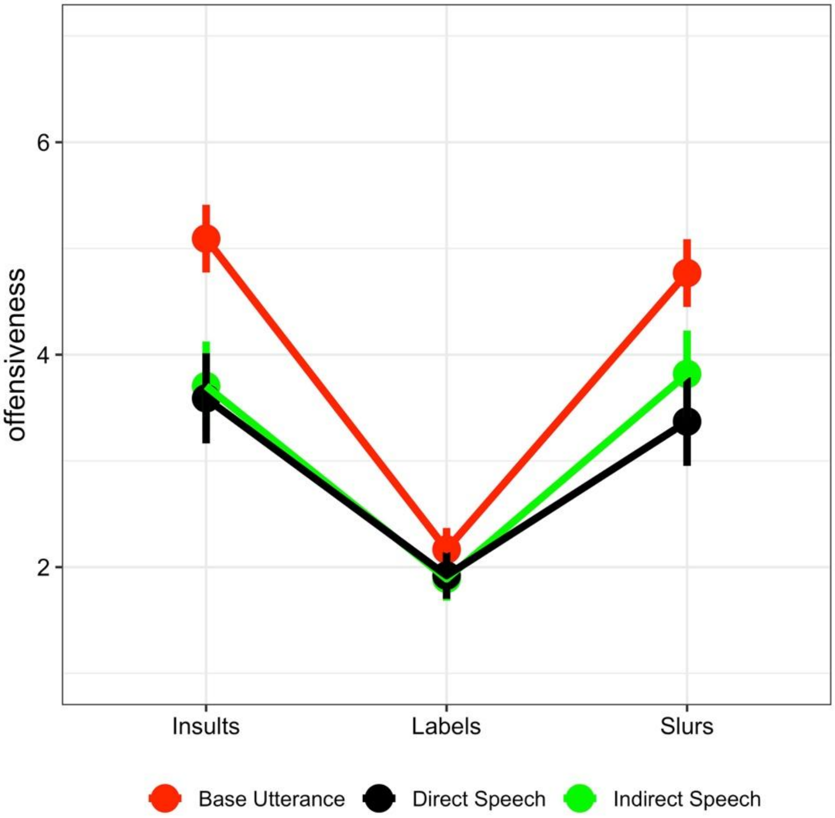

Figure 2. Mean values of offensiveness by condition for Experiment 2. Points represent participants mean scores. Error bars indicate 95% confidence intervals across individuals.

3.3 Discussion

In Study 2 we found that both kinds of reporting strategies—(B) *Direct speech* and (C) *Mixed quotation*—decrease the perceived offensiveness of slurs and insults to a similar extent, with no difference between the two. Both employ quotation marks and display a similar power to decrease the perceived offensiveness of slurs and insults. The fact that no difference was found between *direct speech* (Z: X said: "Y is a P") and *mixed quotation* (Z: X said that Y is "a P") suggests that for both categories, it does not matter how exactly quotation marks are used, that is, whether they concern the whole utterance—"X is a P" —, or the pejorative only—"a P".

4. GENERAL DISCUSSION

In our two studies, we measured the perceived offensiveness of slurs and insults in various kinds of environments, and we found that for all reporting strategies, for both slurs and insults, reports succeed in decreasing their perceived offensiveness, without removing it. Interestingly, for slurs—but not for insults—*indirect speech* got higher offensiveness rates than *direct speech* (Study 1), but no difference emerged between *direct speech* and *mixed quotation* (Study 2). No difference at all was found for insults, for which the only thing that makes a difference in the perceived offensiveness is whether the insult is reported or not. Taken together, our results invite several observations.

First, as expected, all reporting strategies that we have considered decrease the perceived offensiveness of *all* terms: Reporting someone else's speech is an effective tool to dissociate the speaker from the reported content, at least to some extent.

More interestingly, while no difference is found between strategies that employ quotation marks (*direct speech* and *mixed quotation*, Study 2), neither for slurs nor for insults, the perceived offensiveness of reported slurs—but not that of reported insults—decreases more when they are reported in quotation marks, than in *indirect speech*. This asymmetry between slurs and insults suggests that their perceived offensiveness must be sensitive to different factors. Whatever drives the perceived offensiveness of insults, it does not matter *how* the speaker signals that they are reporting someone else's words, as indicated by the absence of any difference between *direct* and *indirect speech* (Study 1), and between *direct speech* and *mixed quotation* (Study 2). In the case of slurs, in contrast, the way in which speakers report the words of others matters. The explicit use of quotation marks—whether they include the whole reported utterance, as in *direct speech*, or only the slur, as in *mixed quotation*—decreases the perceived offensiveness of slurring reports more than *indirect speech*. This asymmetry can be explained by the fact that slurs target social groups, are perceived as a form of hate speech, and thus expose speakers to higher social risks than insults. For slurs, clarifying via quotation marks that the responsibility for the introduction of the pejorative term is the reportee's—rather than the reporter's—matters very much to how offensive the reporting utterance is perceived, much more than it matters for insults. In our studies, the reduction of perceived offensiveness of insults was only determined by the fact that the speaker either used or reported them. On the other hand, the way in which slurs are reported (i.e., whether with or without quotation marks) mattered, and was sufficient to affect their perceived offensiveness.

Finally, another point worth mentioning is that all the reporting strategies that we have considered (those including quotation marks like (B) and (C) and the one employing verba dicendi, like (D)) decreased the perceived offensiveness of *all* the terms, including that (already very low) of non-slurring labels. The perceived offensiveness of such labels is very low in the first place, but significantly higher than that of filler sentences, and it also decreases when they are reported. This finding had been already noted in Cepollaro, Sulpizio and Bianchi (2019) for *indirect speech*. They interpreted the finding that utterances featuring non-slurring labels sound (a tiny bit) offensive as due to the fact that these labels refer to discriminated social groups (such as, e.g., gay people or black people), for which Italian language has slurs. So, ascribing such labels—in the absence of richer context—may produce a (very low) impression of offensiveness, which is further reduced when they are merely reported.

Let us now see how our two studies inform our three initial questions.

The survival question (does the perceived offensiveness of slurs survive when they are reported rather than used?). Our studies provide a mixed answer to the *survival question*: All reporting strategies decrease the perceived offensiveness of slurs, but none of them fully removes it. The same goes for non-slurring insults. This accurately replicates Cepollaro, Sulpizio and Bianchi's (2019) results on Italian slurs in *indirect speech*, and broadens the scope of investigation to two further reporting strategies that employ quotation marks, that is, *direct speech* and *mixed quotation*.

The gradability question (does the perceived offensiveness of slurs come in degrees?). Our results provide a further positive answer to the *gradability question*, in line with the findings of Cepollaro, Sulpizio and Bianchi (2019): The offensiveness of slurs (as well as that of insults) is gradable in the sense that there are intermediate states between the perceived offensiveness of non-reported slurs (our condition A) and its full cancellation (which we did not observe in any condition). Instead, reporting seems able to mitigate, rather than to remove, the perceived offensiveness of these terms. This goes against what certain scholars have argued ("[Slurs'] power to offend … can be not at all mollified by scholarly appeals to use/mention distinctions", Sullivan, 2022, p. 1479), in support of more moderate views (e.g., Bolinger Jorgensen, 2017; Davis & McCready, 2020). If we want to test speakers' intuitions on how the perceived offensiveness of slurs survives across contexts, we cannot expect all-or-nothing results, but rather intermediate states between offensive and non-offensive.

The strategy question (do different discourse reporting strategies affect the perceived offensiveness of slurs in different ways?). Our results provide some initial answers to the *strategy question*, showing that for slurs, the strategies that employ quotation marks are the most effective ones in decreasing the slurs' perceived offensiveness. For slurs, but not for any other category, the strategy employed in reporting another person's speech matters to how offensive the resulting utterance is perceived. In particular, it is crucial that the speaker manifestly signals with quotation marks (whether only on the reported slur or on the entire reported utterance) that they are using someone else's words *verbatim* and that they are not necessarily endorsing the slur themselves.

5. CONCLUSION

Our studies contribute to the literature on slurs and expressives by providing empirical data on Italian slurs and insults. Whether these results can be extended to other languages depends on how much variation slurs and insults display cross-linguistically and cross-culturally. We showed that, at least in Italian, reports succeed in decreasing the perceived offensiveness of slurs and insults, without removing it. We also found an interesting asymmetry between slurs and insults in that for slurs—but not for insults—the strategy employed to report someone else's words matters to how offensive the resulting utterance is perceived. In line with much of the philosophical literature, we explain this asymmetry as follows. Since slurs target social groups and constitute a form of hate speech, they expose speakers to higher social risks. If they do not want to pass as bigots, speakers reporting slurring utterances need to signal their dissociation from the use of a slur as clearly as possible, which they can naturally do by resorting to quotation marks. Once quotation marks are employed, it does not matter much whether they are applied to the whole utterance (*direct speech*) or to a single word only (*mixed quotation*).

The literature on slurs has been growing exponentially over the past decade, resulting in a wide range of theoretical accounts (for a recent survey, see, for instance, Hess, 2022). In theorizing about slurs, authors typically rely on intuitions, which can be conflicting, blurry, and unreliable. Our research provides reliable and robust empirical data on reported slurs—and, for the first time, on *quoted* slurs—against which the existing (and future) theories of slurs and insults may be tested. We believe that the data that we provide remain compatible with many of these theories. Nevertheless, some of the theories may have a harder time than others to account for them. How our findings impinge on the different theories, whether they invalidate any of them or, on the contrary, provide strong evidence for some of them, is an important issue, yet one that we cannot address within the scope of the present paper. Having said that, let us note that, contrary to what one may have thought at a first glance, the fact that slurs are still perceived as offensive even when they are quoted (albeit to a lesser extent than when they are used) is compatible with the predictions of a family of theoretical accounts that take slurs to have derogatory content—a.k.a. *content* theories of slurs.[13] Anderson and Lepore (2013a) contended that such theories must have a hard time explaining why quoted slurs can still be perceived as offensive. However, Rinner and Hieke (2022) have recently proposed a line of explanation, on behalf of such theories, that shows that even if the mechanism of quotation renders the derogatory content inert, a mere mention of a slur can still be perceived as offensive. They argue that, in such cases, offensiveness is diminished with respect to the cases of slurs that are used, which squares well with the findings of our experiments. Their argument is that there is a more general phenomenon for which the content of a given word can have certain effects, also when it is merely quoted. For example, when someone mentions the word “Hitler”, they cause hearers to think of the Nazi dictator and they can provoke emotional responses even if the term is only quoted. They further argue that Anderson and Lepore’s own theory, which takes slurs to be taboo words, has a hard time explaining why quoted slurs are perceived as less offensive than the used ones. Whether our data speak strongly against such theories,

[13] In the family of “content theories”, we find accounts that ascribe some derogatory content to slurs as a lexical category, but differ as to the level at which this content operates—truth-conditional (Hom 2008; Hom & May, 2013), presuppositional (Cepollaro & Stojanovic, 2016; Cepollaro, 2020, among others), conventional implicature (Potts, 2005, 2007), and so forth.

or there is a line of explanation that remains available to them, too, is an issue that must be postponed for future research.

ACKNOWLEDGEMENTS

We take the opportunity to thank two reviewers for their helpful feedback, as well as the audience at the STAL (Slurring Terms Across Languages) Workshop (Paris, June 2023) for their insightful remarks. Claudia Bianchi and Bianca Cepollaro gratefully acknowledge the support of the PRIN project The Mark of Mental (MOM) (Ministero dell'Università e della Ricerca [MUR], 2017P9E9N) and, with Simone Sulpizio, that of the PRIN project EPITHETS - Explaining Pejoratives In THeoretical and Experimental TermS (Ministero dell'Università e della Ricerca [MUR], 2022N87CR9). Isidora Stojanovic acknowledges institutional support from ANR-17-EURE-0017 FrontCog.

REFERENCES

Anderson, L. (2016). When reporting others backfires. In A. Capone, F. Kiefer, and F. Lo Piparo (Eds.), *Indirect reports and pragmatics* (pp. 253-264). Cham: Springer International Publishing.

Anderson, L. & Lepore, E. (2013a). Slurring words. *Nous, 47*(1), 25-48.

Anderson, L. & Lepore, E. (2013b). What did you call me? Slurs as prohibited words. *Analytic Philosophy, 54*(3), 350-363.

Belleri, D. (2020). Slurs: Departures from genuine uses and derogation. *Studies in Logic, Grammar and Rhetoric*, *62*(1), 9-24.

Bolinger Jorgensen, R. (2017). The pragmatics of slurs. *Noûs, 51*(3), 439-462.

Brown, A. (2017). What is hate speech? Part 1: The myth of hate. *Law and Philosophy, 36*, 419-468.

Butler, J. (1997). *Excitable speech: A politics of the performative*. Routledge.

Camp, E. (2018). A dual act analysis of slurs. In D. Sosa (Ed.), *Bad words* (pp. 29-59). Oxford University Press.

Cappelen, H., Lepore, E. & McKeever, M. (2020). Quotation. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (Ed.), URL = <https://plato.stanford.edu/archives/sum2020/entries/quotation/>.

Cepollaro, B. (2020). *Slurs and thick terms.* Rowman & Littlefield.

Cepollaro, B. & Stojanovic, I. (2016). Hybrid evaluatives: In defense of a presuppositional account.

Grazer Philosophische Studien, 93(3), 458–88.

Cepollaro, B., Sulpizio, S. & Bianchi, C. (2019). How bad is to report a slur? An empirical investigation. *Journal of Pragmatics, 146*, 32-42.

Crepaldi, D., Keuleers, E., Mandera, P. & Brysbaert, M. (2013). *SUBTLEX-IT*. Retrieved from. http://crr.ugent.be/subtlex-it/.

Davidson, D. (1979). Quotation. *Theory and Decision, 11*, 27-40. Reprinted in Davidson, D. (1986). *Inquiries into truth and interpretation* (pp. 79-92). Oxford University Press.

Davis, C. & McCready, E., (2020). The instability of slurs. *Grazer Philosophische Studien, 97*(1), 63-85.

Delgado, R. (1982). Words that wound: A tort action for racial insults, epithets, and name-calling. *Harvard Civil Rights-Civil Liberties Law Review, 17*, 133-181.

Eilola, T. M. & Havelka, J., (2010). Affective norms for British English and Finnish nouns. *Behavior Research Methods, 42*, 134-140.

Geisser, S. & Greenhouse, S. (1959). On methods in the analysis of profile data. *Psychometrica, 24*, 95-112.

Gelman, S. A., Ware, E. A. & Kleinberg, F. (2010). Effects of generic language on category content and structure. *Cognitive Psychology*, *61*(3), 273-301.

Harris, J. A. & Potts, C. (2009). Perspective-shifting with appositives and expressives. *Linguistics and Philosophy, 32*(6), 523-552.

Hess, L. (2020). Practices of slur use. *Grazer Philosophische Studien* 97 (1), 86-105.

Hess, L. (2022). Slurs: Semantic and pragmatic theories of meaning. In P. Stalmaszczyk (Ed.), *The Cambridge handbook of philosophy of language* (pp. 450-466). Cambridge University Press.

Hom, C. (2008). The Semantics of Racial Epithets. *Journal of Philosophy, 105*, 416-440.

Hom, C. & May, R. (2013). Moral and semantic innocence. *Analytic Philosophy, 54*(3), 293-313.

Hornsby, J. (2001). Meaning and uselessness: How to think about derogatory words. *Midwest Studies in Philosophy: Figurative Language, 25*, 128-141.

Janschewitz, K. (2008). Taboo, emotionally valenced, and emotionally neutral word norms. *Behavior research methods*, *40*(4), 1065-1074.

Jeshion, R. (2018). Slurs, dehumanization, and the expression of contempt. In D. Sosa (Ed.), *Bad words* (pp. 87-107). Oxford University Press.

Jeshion, R. (2020). Pride and prejudiced. *Grazer Philosophische Studien, 97*(1), 106–137.

Kliegl, R., Grabner, E., Rolfs, M. & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16*, 262-284.

Lasersohn, P. (2007). Expressives, perspective, and presupposition. *Theoretical Linguistics, 33*, 223-230.

Lawrence, M. A. (2016). *Package 'ez'. R package version 4.4-0*.

Lawrence III, C. (1992). Cross burning and the sound of silence: Anti-subordination theory and the first amendment. *Villanova Law Review, 37*, 787-804.

Leiner, D. J. (2019). *SoSci Survey* (Version 3.1.06) [Computer software]. Available at https://www.soscisurvey.de

Leslie, S. J. (2017). The original sin of cognition: Fear, prejudice and generalization. *The Journal of Philosophy*, *114*(8), 393-421.

Maier, E. (2014). Mixed quotation: The grammar of apparently transparent opacity. *Semantics & Pragmatics, 7*(7), 1-67.

Matsuda, M. (1989). Public response to racist speech: Considering the victim's story. *Michigan Law Review, 87*, 2320-2381.

McCready, E. & Davis, C. (2017). An invocational theory of slurs. *Proceedings of LENLS, 14*(1).

Nunberg, G. (2018). The Social Life of Slurs. In D. Fogal, D. Harris & M. Moss (Eds.), *New work on speech act* (pp. 237-295). Oxford University Press.

Panzeri, F. & Carrus, S. (2016). Slurs and negation. *Phenomenology and Mind, 11*, 170-180.

Potts, C. (2005). *The logic of conventional implicatures*. Oxford University Press.

Potts, C. (2007). The expressive dimension. *Theoretical Linguistics*, *33*(2), 165-198.

Predelli, S. (2003). Scare quotes and their relation to other semantic issues. *Linguistics and Philosophy, 26*, 1-28.

Rhodes, M., Leslie, S. J. & Tworek, C.M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*(34), 13526-13531.

Rosenberg, P., Sikström, S. & Garcia, D. (2017). The a (ffective) b (ehavioral) c (ognitive) of taboo words in natural language: The relationship between taboo words' intensity and frequency. *Journal of Language and Social Psychology, 36*(3), 306-320.

Ritchie, K. (2019). Should we use racial and gender generics?. *Thought, 8*(1), 33-41.

Rinner, S. & Hieke, A. (2022). Slurs under quotation. *Philosophical Studies, 179*, 1483-1494.

Saka, P. (2003). Quotational constructions. *Belgian Journal of Linguistics, 17*, 187-212.

Saka, P. (2007). *How to think about meaning*. Springer.

Saul, J. (2017). Are generics especially pernicious?. *Inquiry*, 1-18. 10.1080/0020174X.2017.1285995

Sullivan, A. (2022). Semantic dimensions of slurs. *Philosophia, 50*, 1479-1493.

Sulpizio, S., Toti, M., Del Maschio, N., Costa, A., Fedeli, D., Job, R. & Abutalebi, J. (2019). Are you really cursing? Neural processing of taboo words in native and foreign language. *Brain and language, 194*, 84-92.

Stojnić, U. & Lepore, E. (2022). Inescapable articulations: Vessels of lexical effects. *Noûs, 56* (3), 742-760.

Tirrell, L. (2021). Discursive Eìepidemiology: Two models. *Aristotelian Society Supplementary Volume, 95*(1), 115-142.

Waldron, J. (2010). Dignity and defamation: The visibility of hate. *Harvard Law Review, 123*, 1596-1657.

Williamson, T. (2009). Reference, inference, and the semantics of pejoratives. In J. Almog & P. Leonardi (Eds.), *The philosophy of David Kaplan* (pp. 137-158). Oxford University Press.

Wodak, D., Leslie, S. J. & Rhodes, M. (2015). What a loaded generalization: Generics and social cognition. *Philosophy Compass, 10*(9), 625-635.