Received: 27 August 2020 | Revised: 27 November 2020 | Accepted: 9 March 2021

DOI: 10.1002/tht3.488

ORIGINAL ARTICLE

THOUGHT A JOURNAL OF PHILOSOPHY WILEY

Separating the evaluative from the descriptive: An empirical study of thick concepts

Pascale Willemsen | **Kevin Reuter**

Institute of Philosophy, University of Zurich, Zurich, Switzerland

Correspondence
Pascale Willemsen, University of Zurich, Institute of Philosophy, Zürichbergstrasse 43, CH-8044 Zurich, Switzerland.
Email: pascale.willemsen@uzh.ch

Funding information
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: PCEFP1_181082

Abstract

Thick terms and concepts, such as *honesty* and *cruelty*, are at the heart of a variety of debates in philosophy of language and metaethics. Central to these debates is the question of how the descriptive and evaluative components of thick concepts are related and whether they can be separated from each other. So far, no empirical data on how thick terms are used in ordinary language has been collected to inform these debates. In this paper, we present the first empirical study, designed to investigate whether the evaluative component of thick concepts is communicated as part of the semantic meaning or by means of conversational implicatures. While neither the semantic nor the pragmatic view can fully account for the use of thick terms in ordinary language, our results do favour the semanticist interpretation: the evaluation of a thick concept is only slightly easier to cancel than semantically entailed content. We further discovered a polarity effect, demonstrating that how easily an evaluation can be cancelled depends on whether the thick term is of positive or negative polarity.

KEYWORDS

cancellability, conversational implicature, empirical study, evaluative language, experimental philosophy, moral judgements, polarity effect, thick concepts

Both contributed equally.

1 | SEPARABILITY AND CANCELLABILITY

Philosophers usually distinguish two types of evaluative terms and concepts: thin and thick ones (Eklund, 2011; Väyrynen, 2021). Thin terms evaluate an object as, for instance, "permissible," "wrong," "good," or "blameworthy," yet they do not explicate in what way the object is right or wrong. Thick terms do not merely evaluate, they also provide substantial descriptive information. Typical examples are thick ethical terms and concepts, such as "rude," "reckless," "courageous," or "compassionate." Calling agents courageous evaluates them positively for being willing to take risks—"reckless" also ascribes willingness to take risks yet assigns a negative evaluation to it. While there is widespread consensus that thick terms and concepts somehow unite descriptive and evaluative content, how they actually do that is subject to remarkable disagreement. Some philosophers explain the descriptive richness of thick concepts by arguing that thick terms are basic and inseparable amalgams of description and evaluation (Kirchin, 2010; Putnam, 2002; Roberts, 2011; Williams, 1985). Accordingly, the meanings of thick terms cannot be analysed into a descriptive and evaluative part, but are irreducibly thick.

A majority of philosophers deny this inseparability and claim that thick terms and concepts can be, at least in principle, divided into two distinct components (Blackburn, 1992; Elstein & Hurka, 2009; Hare, 1952). However, these philosophers themselves disagree on how the two components are combined. Semantic separabilists claim that the evaluative component is part of the meaning of a thick term. Thus, similar to the way *unmarried* and *man* can be identified as semantic components of the concept *bachelor*, *willingness to take risks* and *good* can be singled out as separate aspects of the concept of *courage*. The exact analysis is, however, a matter of debate, and, if the semantic view is correct, likely to be more complicated than the case of *bachelor* (see especially Elstein & Hurka, 2009; Kyle, 2019). Pragmatic separabilists claim that the descriptive and the evaluative are connected via pragmatic means, for instance, by conversational implicature (Blackburn, 1992; Hare, 1963; Stevenson, 1938; for discussions of these positions see Eklund, 2011, Kyle, 2013, and Väyrynen, 2013, 2021). Conversational implicatures need to be inferred from what is literally said (Grice, 1989) and come in two variants. Generalised conversational implicatures are communicated across a large variety of contexts. Particularised conversational implicatures, on the other hand, are triggered by the specific communicative circumstances. Väyrynen commits to the view that evaluations are communicated as generalised conversational implicatures (Väyrynen, 2021), but others are less explicit in this respect. The shared assumption is that by saying that an agent is rude, one ascribes some descriptive properties, and one further communicates the implicature that the agent is bad in virtue of having these properties.

Arguments in favour of either the semantic or the pragmatic separabilist position heavily rely on linguistic intuitions about how thick terms expressing thick concepts can be used. Fortunately, experimental linguistics provides the means to test the intuition that the evaluative component is merely communicated via conversational implicatures, namely the cancellability test (Grice, 1989, but see Blome-Tillmann, 2008; Sullivan, 2017; Zakkou, 2018 for discussions of the test's limitations). If the pragmatists are correct and the evaluative aspect is only conversationally implicated, cancelling the evaluation should not lead to a contradiction. Take, for instance, the sentence "There is the door." This statement not only communicates the location of a door, but in some contexts carries the particularised conversational implicature that the addressee is asked to leave the room. And yet, saying "There is the door, but I am not saying you should leave" does not yield a contradiction. Generalised conversational implicatures work

in a similar way but depend less on the specific context. Saying "I ate some cookies" reliably communicates the implicature that the speaker ate some, but not all the cookies. A statement like "I ate some of the cookies, but by that I am not saying that I did not eat all of them" is nevertheless non-contradictory. If the evaluation of a thick term is conversationally implicated by a thick term, cancelling the evaluation should be equally non-contradictory. Therefore, a speaker who utters "What Tom did was rude, but by that I am not saying something negative about Tom" makes a felicitous, non-contradictory statement. If we were to find empirical evidence that cancelling the evaluation is non-contradictory, this would count as evidence for the pragmatist position which treats the evaluation as a conversational implicature.

In contrast, semantic separabilists claim that the evaluative component cannot be cancelled in the way suggested above. The evaluative component is semantically entailed by the use of a thick term, just as "Tom is unmarried" is entailed by "Tom is a bachelor." Accordingly, a speaker who utters "Tom is a bachelor, but by that I am not saying he is unmarried" contradicts herself. If the evaluative part of a thick term is also semantically entailed, cancelling the evaluation should be similarly contradictory. Thus, a person who says "What Tom did was rude, but by that I am not saying something negative about Tom" makes an infelicitous statement. With this well-established test at our disposal, we designed and pre-registered an experiment aiming to decide between advocates of the semantic view and adherents of pragmatic separability.

At this point, we would like to make three remarks. First, this paper assumes that separabilism is a theoretically plausible account that is worthy of being tested empirically. We do not, however, commit ourselves to the correctness of separabilism and the falsity of inseparabilism. Second, while we do not focus on inseparabilism, our results still indirectly inform this position as well. Any result in favour of a semanticist account will be compatible with inseparabilist accounts as well. In contrast, low contradiction ratings would not only disprove the semanticists, any inseparabilist account would not even get off the ground. Third, the distinction between the pragmatists and the semanticists does not adequately reflect the variety of separabilist accounts on the table. Specifically, pragmatists do not need to argue that the evaluative part is communicated via conversational implicature, which is easy to cancel without creating a contradiction. Instead, it might be argued that the evaluation is conventionally implicated (Zakkou, n.d.) or presupposed (Cepollaro, 2020; Cepollaro & Stojanovic, 2016), and cannot be easily cancelled. Consequently, if we find that the evaluation cannot be cancelled, this evidence does not prove *all* pragmatists wrong.

2 | EMPIRICAL STUDY

To the best of our knowledge, no empirical results exist on whether the evaluative component of a thick term is connected to the descriptive component through conversational implicature. The aim of this study was to provide this long overdue evidence. We presented 868 participants with sentences in which a particularised or a generalised conversational implicature, a semantic entailment, or the evaluation of a thick term was first communicated and then cancelled. We then asked participants to what extent they believed that the speaker contradicted herself. We predicted that for both particularised and generalised conversational implicatures, cancelling the implicated meaning would be possible without creating a contradiction. For semantic entailments, cancelling should not be possible and result in high contradiction ratings. For thick terms, we hypothesised that if the pragmatists were correct, cancelling the evaluation should provide contradiction ratings similar to those for conversational implicatures. If the

semanticists were correct though, contradiction ratings would resemble those of semantic entailments. The experimental design, predictions, and statistical models were pre-registered with the Open Science Framework (osf.io/xew6d).[1]

Participants were recruited via Prolific Academics and completed an online survey implemented in Qualtrics. All participants were required to be at least 18 years old, English native speakers, and to have an approval rate of previous studies on the platform of at least 80%. Before engaging with the actual experimental stimuli, all participants were provided with a training in which we familiarised them with the term 'contradict'. Participants then answered two test questions, and we only included participants who answered at least one of them correctly. We excluded 89 participants for failing the training round.[2] A total of 779 participants were included in the analysis (36.87% male, 62.35% female, 0.78% non-binary; $M_{age} = 34.46$).

We implemented a 7×1 between-subject design with the independent variable *Condition* (Thick Concepts Behaviour Positive [short: TCBehaviourPos], Thick Concepts Behaviour Negative [TCBehaviourNeg], Thick Concepts Character Positive [TCCharacterPos], Thick Concepts Character Negative [TCCharacterNeg], Semantic Entailment [SE], Particularised Conversational Implicature [PCI], and Generalised Conversational Implicature [GCI]) and the dependent variable *Contradiction*. As stimuli, we used:

- six negative thick concepts: cowardly, cruel, manipulative, rude, selfish, vicious
- six positive thick concepts: compassionate, courageous, friendly, generous, honest, virtuous
- four particularised conversational implicatures: chocolate, dark, door, hungry
- four generalised conversational implicatures: and, man, some, tried
- four semantic entailments: couch, lake, run, widow

Thick concepts were selected based on seven criteria[3] First, we decided to only test adjectives that can be used to describe a person's behaviour or character and fit in a sentence like "What Amy did was X" or "Amy is X." Second, we chose items which are frequently discussed as examples of thick concepts in the philosophical literature. Third, we selected thick terms that are frequently used in ordinary language. Fourth, we created pairs of items which constitute opposites of one another. Fifth, these opposites do not share the same word stem, as would be the case with "honest" and "dishonest," to avoid possible confounding effects of such constructions. Sixth, to ensure that all items are clearly positive and clearly negative, we determined each item's sentiment value and paired items with similar evaluative intensity. Finally, we excluded terms that bear a risk of communicating an objectionable and thus not widely shared evaluation, such as terms connected to religiosity and sexual morale (for discussions of objectionable thick concepts, see Väyrynen, 2009, 2013).[4]

Here are some concrete examples of the sentences we used:

- negative/positive thick Behaviour: "Amy's behavior last week was rude/friendly, but by that I am not saying something negative/positive about Amy's behavior that day."
- negative/positive thick Character: "Amy is rude/friendly, but by that I am not saying something negative/positive about Amy."
- particularised conversational implicature: "This chocolate is good value-for-money, but by that I am not saying that we should buy it."
- generalised conversational implicature: "Zoe ate some of the cookies, but by that I am not saying that she did not eat all of them."

- semantic entailment: "This is a couch, but by that I am not saying that this is a piece of furniture."

Participants then answered the question "Does Sally contradict herself?" on a scale from "1= definitely not" to "9 = definitely yes."

Participants in the four thick concept conditions read the stimuli for all six items in their condition. Participants in PCI and GCI read the stimuli for all four particularised or generalised conversational implicatures, and participants in SE read the stimuli for all four semantic entailments. All stimuli were presented in randomised order. As philosophers in the debate usually make the tacit assumption that thick concepts form a uniform class of concepts, we do not believe they would predict an effect of polarity for thick concepts. For this reason, we collapsed items of positive and negative polarity for some statistical analyses.

We conducted a global 7 × 1 ANOVA with *Condition* as a between-subject factor and *Contradiction* as the dependent variable. The results for the seven conditions are depicted in Figure 1 and Table 1. Appendix provides the statistics for each item we tested. The analysis revealed a significant effect of *Condition*, $F(5, 11{,}551) = 268.6$, $p < .001$. In accordance with our pre-registered hypotheses, we conducted two planned contrasts, namely for SE and PCI and for SE and GCI. The mean value for SE (7.33, SD = 2.72) was significantly higher compared to PCI (2.70. SD = 2.51), $t(885.54) = -26.46$, $p < .001$, and also significantly higher than the mean value for GCI (3.04), $t(873.44) = -22.64$, $p < .001$. Our baseline conditions thus worked as expected.

The philosophically motivated predictions we made about thick concepts do no assume an effect of polarity. We therefore collapsed positive and negative thick terms and ran additional planned contrasts. We compared TCBehaviour (6.64, SD = 2.68) with SE and found a significant difference, $t(756.28) = -4.73$, $p < .001$. We ran the same test for TCCharacter (6.72,

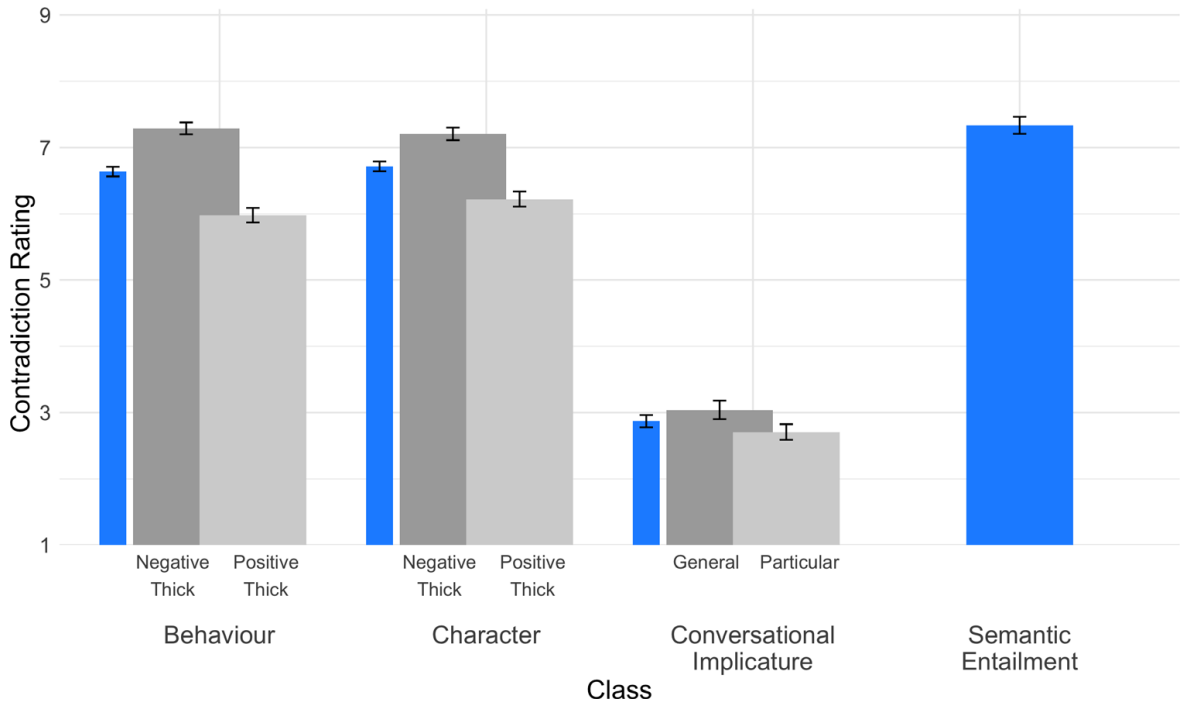

FIGURE 1 Average contradiction ratings of the various conditions. The blue bars show the mean values of the collapsed data. The grey bars the means for the unpooled data. The error bars indicate the standard error around the means

TABLE 1 Summary statistics of the main conditions

Nested group	Mean	Median	SE	SD
Behaviour: Negative thick	7.29	8	0.09	2.35
Behaviour: Positive thick	5.98	7	0.11	2.83
Character: Negative thick	7.21	8	0.10	2.48
Character: Positive thick	6.22	7	0.11	2.93
Conversational implicature: General	3.04	1	0.14	2.92
Conversational implicature: Particular	2.70	1	0.12	2.51
Semantic entailment	7.33	9	0.13	2.72

SD = 2.76) and SE and found a significant difference as well, $t(777.77) = -4.16$, $p < 0.001$. Based on these results, we can conclude that in both embeddings, contradiction ratings for thick concepts are significantly different from those for semantic entailments. The prediction of the Semantic View was therefore not met.

To test the Pragmatic View, we compared both PCI and GCI with TCBehaviour and TCCharacter. All four comparisons revealed significant differences (PCI vs. TCBehaviour: $t(801.34) = 28.17$, $p < .001$; GCI vs. TCBehaviour: $t(687.63) = 22.82$, $p < .001$; PCI vs. TCCharacter: $t(825.97) = 28.48$, $p < .001$; GCI vs. TCCharacter: $t(705.38) = 23.15$, $p < .001$). The prediction of the Pragmatic View was also not met.

Going beyond the philosophical literature and in line with our pre-registered hypotheses, we tested whether the polarity of thick concepts has an effect on contradiction ratings. We predicted that in both embeddings, contradiction ratings for negative terms would be significantly higher than ratings for positive terms. This prediction was confirmed for both the Behaviour $M_{\text{negative}} = 7.29$, $M_{\text{positive}} = 5.98$; $t(1297.8) = 9.26$, $p < .001$) and the Character condition ($M_{\text{negative}} = 7.21$, $M_{\text{positive}} = 6.22$, $t(1297) = 6.64$, $p < .001$).[5] This polarity effect was observed across all paired adjectives that we selected for our study, that is, all negative thick adjectives received descriptively higher contradiction ratings than their positive counterparts, for example, "cowardly" versus "courageous," or "rude" versus "friendly" (see also Appendix for the summary statistics for each term).

3 | DISCUSSION

Is the evaluative aspect of a thick term conveyed by means of conversational implicature—as many pragmatists argue—, or is the evaluative component semantically entailed? In this paper, we presented the results of an empirical study on thick concepts focusing both on the relation between the evaluative and descriptive aspects of thick concepts. We distinguished two views on how evaluative and descriptive content are combined, namely the Inseparabilist and the Separabilist position. According to Separabilists, evaluation and description are distinct components which can be distinguished from one another. Among Separabilists, scholars disagree on whether these distinct components are semantically or pragmatically related. The cancellability test was used to address this disagreement.

Neither the predictions of the pragmatist view nor the semanticist view were met, albeit to different degree. Against the pragmatists' prediction, the evaluation of a thick concept was

significantly harder to cancel compared to conversationally implicated content. This effect maintained for two different embeddings of thick terms, and also when thick concepts were compared to generalised and particularised conversational implicatures. Challenging the semanticist, the evaluation of thick concepts was significantly easier to cancel compared to semantically entailed content.

That said, the degree to which the mean values for the thick concepts conditions differed from the mean values for semantic entailment and conversational implicature, were highly dissimilar. For both TCBehaviour and TCCharacter, the mean values were only slightly lower compared to the average result for the semantic entailment condition (SE vs. TCBehaviour: $\Delta = 0.69$, SE vs. TCCharacter: $\Delta = 0.61$). In contrast, those same mean values were much higher than the average results for particularised conversational implicatures (PCI vs. TCBehaviour: $\Delta = 3.94$, PCI vs. TCCharacter: $\Delta = 4.02$) and generalised conversational implicatures (GCI vs. TCBehaviour: $\Delta = 3.60$, GCI vs. TCCharacter: $\Delta = 3.68$). Semanticists might not be too worried by the observed differences we found. And indeed, since contradiction ratings were only slightly lower from those for semantic entailments, our data does favour the semanticist interpretation. Of course, an explanation should be provided as to why cancellation of the evaluative component is that bit easier compared to standard cases of semantic entailment. Some insights into what might be going on can be provided by looking at the differences between positive and negative thick concepts, to which we now turn.

Our study revealed a polarity effect on contradiction ratings. For positive thick terms, contradiction ratings were significantly lower compared to negative thick terms as well as semantic entailments. This polarity effect is hitherto unknown and has not been predicted by any of the various accounts of thick concepts. In fact, the effect challenges the tacit assumption that thick terms and concepts form a homogenous group for which we can ask broad questions about separability and how evaluation and description are connected. If we only look at the average results for negative thick concepts, there is no significant difference compared to the mean values for the semantic entailment condition.[6] Thus, it might be proposed that we need two separate accounts of thick concepts, one for positive and one for negative concepts. Alternatively, the semanticist might want to argue that our data on negative thick concepts correctly reflects the semantic view, and our results on positive terms are confounded by a further factor.[7] The plausibility of those two proposals will depend on which account best explains the polarity effect. While we do not yet have strong evidence in favour of any one of these accounts, here are two explanations that we consider plausible:

A first explanation focuses on differences in the availability of counterexamples to the usually communicated evaluation. One might claim that when thinking about honesty and courage, we think of cases in which an agent is being too honest or too courageous or in which they are honest or courageous, yet for the wrong reasons. Moreover, virtues such as honesty usually interact with other virtues, for example politeness, respect towards others, etc.[8] A situation might render honesty the wrong course of action because politeness and respect are more important virtues in that context.[9] Consequently, when using positive thick terms like "honest," we quite readily come up with examples in which being honest is not such a good thing but has (at least partially) turned into something negative. Thinking about such cases then provides a way of making sense of an otherwise contradictory statement. In contrast, for negative thick terms, such counterexamples are hardly available. It is difficult to create an example in which an agent's behaviour is too cruel, and therefore good, a case in which one is cruel for the right reasons, or in which other vices interact with cruelty, making cruelty the right course of action.

We believe that this explanation of the polarity effect is very powerful and deserves further research. However, the data we collected provides a direct challenge to it. The polarity effect occurred not only for uses of thick terms in which they were attributed to a person's behaviour, but also when being attributed to a person's character. While it is relatively easy to think of individual scenarios in which honest behaviour might not be a good thing, examples are less forthcoming that deny the goodness of honesty as a character trait. Honesty, understood as a disposition to bring about a certain type of behaviour, will result more often in good than in bad behaviour. Therefore, honesty as a character trait should be evaluated positively; and this evaluation should be difficult to cancel. Our results are not in line with this reasoning.[10]

Second, one might wonder whether the polarity effect can be explained by different social norms that guide evaluative language. Uttering a positive thick term without the intention to commit to a positive evaluation seems relatively harmless. Being misunderstood in cases of negative thick terms has a potentially greater impact. If mistaken, a speaker communicates a negative evaluation they initially did not want to commit to. Since negative evaluations harm others by diminishing their social status and reputation, people might well be less willing to accept a cancellation of a negative evaluation. This is not to say that there is a strong connection between censoring a speaker and rating her high on contradictoriness, but only that contradiction ratings might serve as proxy for a tendency to censor such behaviour. While so far, we have no empirical evidence speaking either in favour of or against this hypothesis, given our knowledge about the effects of norm violations on a variety of non-normative concepts, it seems quite plausible to assume that norm-violations can affect contradiction ratings differently. Over the past 20 years, a growing body of empirical evidence suggests that moral valence has a significant effect on judgements about causation (Sytsma et al., 2019, for an overview see Willemsen & Kirfel, 2019), intentionality (Knobe, 2003), knowledge (Beebe & Buckwalter, 2010), just to name a few.

If that were the case, then one might posit that the recorded effect of people's contradiction ratings for negative thick concepts is increased due to such a negativity bias. This bias will arguably be weakened when it comes to positive thick terms, suggesting that the results we collected for positive thick concepts give us a less distorted view of the relation between the descriptive and evaluative components of thick concepts. On the other hand, it is likely that the use of positive thick terms is also subject to social norms, even if less stringent. Disentangling the effects of social norms from the mere linguistic aspects will be a serious challenge yet to be overcome.[11]

ORCID

Pascale Willemsen https://orcid.org/0000-0002-4563-1397
Kevin Reuter https://orcid.org/0000-0003-2404-1619

ENDNOTES

[1] The experiment initially submitted was based on Willemsen & Reuter, 2020. We thank three anonymous reviewers for their critical feedback which helped us to significantly improve the initial design.

[2] The instructions that were given to participants, as well as all stimuli sentences are available in this online repository. https://mfr.de-1.osf.io/render?url=https://osf.io/n973r/?direct%26mode=render%26action=download%26mode=render

[3] All selection criteria are elaborated on in more detail in the preregistration.

[4] We would like to emphasise that in this paper, we only tested thick *ethical* concepts. Therefore, our discussion of the semanticist and pragmatist predictions are limited to our findings in this very specific domain. It might be suspected that thick aesthetic or epistemic concepts work differently in the cancellation test, thereby providing different support for semanticist and pragmatist positions. We aim to test this in future studies.

[5] As an exploratory analysis, we analysed whether there was an interaction between Polarity and Embedding, such that the polarity effect is of different size in the two embeddings. The interaction was not significant ($p = .113$).

[6] An exploratory analysis of the results for TCCharacterNeg and SE revealed no significant difference, $p = 0.424$. A similar outcome was obtained when we compared TCBehaviourNeg with SE, $p = 0.771$.

[7] Of course, pragmatists will turn the argument around and claim that the results for positive thick terms are less distorted. However, the rather large difference in contradiction ratings compared to conversational implicatures leaves a massive gap in that argument. Our results might thus be more encouraging to other variants of the pragmatist account like those arguing that evaluation is conventionally implicated or presupposed.

[8] We would like to thank an anonymous reviewer of this journal for this idea.

[9] Note that these considerations might be spelled out as three different and independent versions of the availability of counterexamples explanation, and there might be even more. Each version should be worked out in more detail, including the empirical predictions they make.

[10] We did collect data on participants' response times to the test questions that indicate that the availability of counterexamples might have played a role in people's deliberations. We compared the duration of the time spent to answer a single test question between those participants who received positive and those who received negative thick terms (excluding participants with response times two SDs different from the means). On average, participants who were presented with a positive thick term took 7.49 s (SD = 5.10) to rate the contradictoriness of the statement in the Behaviour condition and 5.37 s (SD = 3.91) in the Character condition. In contrast, participants spent less time (6.60 s [SD = 4.97] in the Behaviour condition and 4.07 s [SD = 3.37] in the Character condition), when they were given the negative thick terms. These differences were not very large but significant: Behaviour: $t(1310) = 3.213$, $p = .001$; Character: $t(1303) = 2.011$, $p = .045$.

[11] We would like to thank three anonymous reviewers and the editor of this journal for their invaluable feedback on our paper. We are grateful to Lucien Baumgartner, Charles Dordjevic, Catherine Herfeld, Guido Loehr, Judith Martens, and Julia Zakkou for their critical and very constructive comments on earlier drafts of this paper. This project was presented in the online colloquium of the Katedra filosofie a dějin přírodních věd at Charles University Prague, at the annual meeting of the Cognitive Science Society 2020, at the Group for Empirical Approaches to Morality and Society (New York), at the workshop on Empirical Moral Psychology at the University of Salzburg, and at the Experimental Philosophy of Language and Metaethics workshop at the University of Bochum. We thank all participants for their questions and comments. We are also grateful to Lucien Baumgartner for creating Figure 1. This research was funded by the Swiss National Science Foundation (grant number PCEFP1 181082).

REFERENCES

Beebe, J., & Buckwalter, W. (2010). The epistemic side-effect-effect. *Mind & Language*, *25*(4), 474–498.

Blackburn, S. (1992). Through thick and thin. *Proceedings of the Aristotelian Society*, *66*, 284–299.

Blome-Tillmann, M. (2008). Conversational implicature and the cancellability test. *Analysis*, *68*(2), 156–160.

Cepollaro, B. (2020). *Slurs and thick terms. when language encodes values*. Roman & Littlefield.

Cepollaro, B., & Stojanovic, I. (2016). Hybrid evaluatives: In defense of a presuppositional account. *Grazer Philosophische Studien*, *93*(3), 458–488.

Eklund, M. (2011). What are thick concepts? *Canadian Journal of Philosophy*, *41*(1), 25–49.

Elstein, D., & Hurka, T. (2009). From thick to thin: Two moral reduction plans. *Canadian Journal of Philosophy*, *39*(4), 515–536.

Grice, P. (Ed.) (1989). Logic and conversation. In *Studies in the way of words* (pp. 22–40). Harvard University Press.

Hare, R. (1952). *The language of morals*. Clarendon Press.
Hare, R. M. (1963). Freedom and Reason. *Reprint. Clarendon Paperbacks*, Oxford: Clarendon Press.
Kirchin, S. (2010). The shapelessness hypothesis. *Philosophers' Imprint*, *10*(4), 1–28.
Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(279), 190–194.
Kyle, B. (2013). How are thick terms evaluative? *Philosophers' Imprint*, *13*(1), 1–20.
Kyle, B. (2019). The expansion view of thick concepts. *Noûs*, *54*(4), 914–944.
Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays*. Harvard University Press.
Roberts, D. (2011). Shapelessness and the thick. *Ethics*, *121*(3), 489–520.
Stevenson, C. L. (1938). Persuasive definitions. *Mind*, *47*(187), 331–350.
Sullivan, A. (2017). Evaluating the cancellability test. *Journal of Pragmatics*, *121*, 162–174.
Sytsma, J., Bluhm, R., Willemsen, P., & Reuter, K. (2019). Causal attributions and corpus linguistics. In E. Fischer & M. Curtis (Eds.), *Methodological advances in experimental philosophy*. Bloomsbury.
Väyrynen, P. (2009). Objectionable thick concepts in denials. *Philosophical Perspectives*, *23*(1), 439–469.
Väyrynen, P. (2013). *The lewd, the rude and the nasty*. Oxford University Press.
Väyrynen, P. (2021). Thick Ethical Concepts. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/spr2021/entries/thick-ethical-concepts
Willemsen, P., & Kirfel, L. (2019). Recent empirical work on the relationship between causal judgments and norms. *Philosophy Compass*, *14*(1), e12562.
Willemsen, P., & Reuter, K. (2020). Separability and the effect of valence. In M. Denison & A. Xu (Eds.), *Proceedings of the 42th annual conference of the cognitive science society 2020* (pp. 794–800). Cognitive Science Society. https://cognitivesciencesociety.org/wp-content/uploads/2020/07/cogsci20_proceedings_final.pdf
Williams, B. (1985). *Ethics and the limits of philosophy*. Harvard University Press.
Zakkou, J. (2018). The cancellability test for conversational Implicatures. *Philosophy Compass*, *13*(12), e12552.
Zakkou, J. (n.d.). Conventional Evaluativity.

How to cite this article: Willemsen P, Reuter K. Separating the evaluative from the descriptive: An empirical study of thick concepts. *Thought: A Journal of Philosophy*. 2021; 1–12. https://doi.org/10.1002/tht3.488

APPENDIX

A SUMMARY STATISTICS FOR TESTED ITEMS

Class	Term	Mean	Median	SE	SD
Behaviour	Cowardly	6.45	7	0.25	2.69
Behaviour	Cruel	8.05	9	0.15	1.63
Behaviour	Manipulative	6.93	8	0.24	2.55
Behaviour	Rude	7.65	8	0.19	2.03
Behaviour	Selfish	6.80	8	0.24	2.52
Behaviour	Vicious	7.85	9	0.20	2.08
Behaviour	Compassionate	6.60	7	0.24	2.56
Behaviour	Courageous	5.38	6	0.28	3.01
Behaviour	Friendly	6.44	7	0.24	2.59
Behaviour	Generous	6.23	7	0.25	2.69
Behaviour	Honest	5.12	5	0.29	3.07
Behaviour	Virtuous	6.09	7	0.26	2.77
Character	Cowardly	6.35	7	0.25	2.66
Character	Cruel	7.88	9	0.20	2.11
Character	Manipulative	6.99	8	0.23	2.47
Character	Rude	7.59	9	0.22	2.35
Character	Selfish	7.11	8	0.23	2.43
Character	Vicious	7.32	9	0.25	2.59
Character	Compassionate	6.85	8	0.26	2.76
Character	Courageous	6.00	7	0.27	2.87
Character	Friendly	6.44	8	0.27	2.87
Character	Generous	6.67	8	0.27	2.83
Character	Honest	5.57	7	0.30	3.13
Character	Virtuous	5.80	6	0.28	2.94
Part. Conv. Implicature	Chocolate	1.93	1	0.17	1.83
Part. Conv. Implicature	Dark	2.51	1	0.23	2.45
Part. Conv. Implicature	Door	2.05	1	0.18	1.87
Part. Conv. Implicature	Hungry	4.33	4	0.28	2.95
Gen. Conv. Implicature	And	3.75	2	0.32	3.33
Gen. Conv. Implicature	Man	1.50	1	0.12	1.30
Gen. Conv. Implicature	Some	4.34	3	0.32	3.31
Gen. Conv. Implicature	Tried	2.56	1	0.23	2.40
Semantic Entailment	Couch	8.07	9	0.21	2.22
Semantic Entailment	Lake	6.10	7	0.29	3.05

(Continues)

Class	Term	Mean	Median	SE	SD
Semantic Entailment	Run	7.03	9	0.27	2.91
Semantic Entailment	Widow	8.14	9	0.19	2.04