Nat Lang Semantics
DOI 10.1007/s11050-016-9130-7

On how verification tasks are related to verification procedures: a reply to Kotek et al.

Tim Hunter[1] · **Jeffrey Lidz**[2] · **Darko Odic**[3] · **Alexis Wellwood**[4]

Abstract Kotek et al. (Nat Lang Semant 23: 119–156, 2015) argue on the basis of novel experimental evidence that sentences like 'Most of the dots are blue' are ambiguous, i.e. have two distinct truth conditions. Kotek et al. furthermore suggest that when their results are taken together with those of earlier work by Lidz et al. (Nat Lang Semant 19: 227–256, 2011), the overall picture that emerges casts doubt on the conclusions that Lidz et al. drew from their earlier results. We disagree with this characterization of the relationship between the two studies. Our main aim in this reply is to clarify the relationship as we see it. In our view, Kotek et al.'s central claims are simply logically independent of those of Lidz et al.: the former concern which truth condition(s) a certain kind of sentence has, while the latter concern the procedures that speakers choose for the purposes of determining whether a particular truth condition is satisfied in various scenes. The appearance of a conflict between the two studies stems from inattention to the distinction between questions about truth conditions and questions about verification procedures.

Keywords Verification · Experimental semantics · Cognition · Most · Quantification

✉ Tim Hunter
timhunter@ucla.edu

[1] Department of Linguistics, University of California, Los Angeles, 3125 Campbell Hall, UCLA, Los Angeles, CA 90095-1543, USA

[2] Department of Linguistics, University of Maryland, 1401 Marie Mount Hall, College Park, MD 20742, USA

[3] Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, BC V6T 1Z4, Canada

[4] Department of Linguistics, Northwestern University, 2016 Sheridan Rd, Evanston, IL 60208, USA

Published online: 18 November 2016

1 Introduction

Kotek et al. (2015) (henceforth KSH) argue on the basis of novel experimental evidence that sentences like (1) exhibit an ambiguity between two readings. The first reading is the familiar one, under which the sentence is true iff there are more blue dots than non-blue dots; KSH call this the *proportional* reading. KSH argue that this sentence has a second, *superlative* reading, under which the sentence is true iff there are more blue dots than dots of any other color.

(1) Most of the dots are blue.

KSH's evidence for this superlative reading comes from a picture verification task. The crucial finding, discussed in more detail below, is that participants' true/false judgements for the sentence in (1) are sensitive to manipulations of the makeup of the non-blue dots. In particular, the rate of 'yes' responses on trials with two non-blue colors (yellow and red) differed from that on trials with only one non-blue color (yellow). (In addition—but less crucially for present purposes—among those trials with two non-blue colors present, the rate of 'yes' responses on trials with an unbalanced blue:yellow:red ratio such as 10:9:1 differed from the rate on trials with a more balanced ratio of 10:6:4.)

According to KSH, this crucial finding "appears to be in conflict with" (p. 150) results reported in the earlier work of Lidz et al. (2011). Lidz et al. also conducted an experiment using a picture verification task based on the sentence in (1), and found no difference in participants' rate of 'yes' responses across trials with one, two, three, or four non-blue colors present. As KSH note, however, this conflict is only illusory, because there are important differences between the two picture-verification tasks. Lidz et al.'s task presented the stimuli for only a very brief period of time (150 ms), whereas KSH's task allowed participants to look at the stimuli for as long as desired. We agree with KSH that the differences between the tasks can explain the "differences" between the findings. But we disagree with KSH's claim that the integrated picture one arrives at after taking the task differences into account "casts doubt on the generalizability of the results and on the robustness of the conclusions drawn from [Lidz et al.]" (p. 151).

The task differences between KSH and Lidz et al. (2011) reflected the fact that the two studies had very different aims: KSH addressed a question about which truth condition(s) sentence (1) has, whereas Lidz et al. addressed a question about which procedure participants use to determine whether this sentence's proportional truth condition was satisfied in a given scene. These differences are possibly obscured by the differing ways in which the two papers use the term "verification" (we return to this point in Sect. 4). The combined findings of the two studies are compatible with the conjunction of KSH's claims about truth conditions[1] and Lidz et al.'s claims about procedures. This is not a situation where two experiments addressing the same

[1] We are actually somewhat sceptical of KSH's claims about the superlative truth condition, for reasons we return to in Sect. 5. But for expository purposes we will grant them this point for most of what follows, in order to show how the conclusions of Lidz et al. remain valid *even if* KSH's claims about the superlative truth condition are taken to be correct.

question reached different findings, one of which fully answers the question and the other of which, due to irrelevant task effects, presents a distorted or incomplete view. Rather, it is a situation where two distinct experiments addressed two distinct questions.

We will review the Lidz et al. (2011) study in Sect. 2, and then, against the backdrop of the goals and claims of that earlier study, discuss KSH in Sect. 3. In Sect. 4 we explain why we take the conflicts between the two studies to be only apparent.

2 Lidz et al.'s investigation of verification procedures

The questions addressed by Lidz et al. (2011) might be described as "sub-truth-conditional": they are questions that remain unanswered after the truth conditions of sentences are specified. Lidz et al.'s main claim is that their findings are best explained by supposing that, contrary to common idealizing assumptions, a comprehensive theory of the meaning of a sentence like (1) will need to take into account not only facts about truth conditions but also (at least some) facts about these finer-grained, sub-truth-conditional issues.

A truth condition amounts to a function from contexts (possibly formalized as worlds, or models—we will use the relatively non-committal term "contexts") to truth values: the familiar truth condition most often associated with (1), for example, can be identified with the function that maps all those contexts where there are more blue dots than non-blue dots to True, and maps all other contexts to False. A standard and reasonable step in describing the semantics of a certain sentence S is to identify a truth condition (or, in the case of ambiguity, truth conditions) in accord with which a speaker judges S to be true or false—i.e., to identify a truth condition that maps to True exactly those contexts in which a speaker judges S to be true.

Identifying a given truth condition, however, leaves open the question of how the speaker goes about determining whether, in a given scene, the condition is met; in other words, what computations the speaker applies to the available information about the context in order to arrive at a true/false judgement. We will call an answer to this "how" question a verification procedure; specifically, we will say that a verification procedure P *implements*[2] a function (or truth condition) F iff, for any input x, running P on x produces the result F(x). So the important point is that truth conditions stand in a one-to-many relationship with verification procedures: for any given truth condition, there are many distinct verification procedures that implement it (see Marr 1982; Pietroski et al. 2009).[3]

[2] Our use of the term 'implement' here still abstracts away from questions about the physical realization of a computation (Marr's level three); instead it refers to the relationship between functions (Marr's level one) and algorithms or procedures (Marr's level two).

[3] Note that it makes no sense to talk about, for example, "the function that compares an approximation of the numerosity of the blue dots with an approximation of that of the yellow dots, and returns True if the former exceeds the latter and False otherwise". A function is entirely characterized by its input-output pairings. In particular, a function mapping contexts to truth values is entirely characterized by the set of contexts that it maps to True. See Gallistel and King (2009, Chap. 3).

The experiment reported by Lidz et al. was designed to probe the question of which verification procedure(s), of the many conceivable ones that implement the familiar truth condition expressed in (2), speakers use to arrive at a true/false response to the target statement in (1).

(2) $|\text{DOT} \cap \text{BLUE}| > |\text{DOT} \setminus \text{BLUE}|$

(This is what KSH call the proportional truth condition. The possibility that 'most' has the superlative truth condition was not discussed in Lidz et al.'s paper. We return to this point in Sect. 4.)

The experimental task was carefully designed to ensure that—crucially—*two distinct verification procedures* were both available (in certain critical trials). The nature of the task did, of course, rule out many, many other verification procedures that one might in principle use to arrive at true/false responses to (1) in other circumstances. But the task was chosen so as to not rule out so many verification procedures that only one would be left as a viable option. Despite this, the results indicated that participants consistently used only one of the two verification procedures that were left as viable options, even in situations where the other viable procedure would have yielded more accurate responses. This was the crucial finding that Lidz et al. aimed to draw attention to, for this is the point which suggests that a truth condition does not exhaustively characterize the meaning of the target sentence. Rather, there must in addition be some aspect of the sentence's meaning that favors one procedure over another, despite their truth conditional equivalence.

The task that produced this result worked as follows. On each trial, participants saw an array of colored dots like the ones in Fig. 1 displayed on a computer screen for 150 ms—too briefly to allow precise counting—and were asked to answer the question "Are most of the dots blue?". Participants answered 'yes' or 'no' by pressing a button on a keyboard. One dimension in which the displays varied was the ratio of blue to non-blue dots (1:2, 2:3, 3:4, 5:6, 7:8, and the inverses of these ratios). The more significant manipulation was variation in the number of non-blue colors present: this number varied from one (as in Fig. 1a) to four (as in Fig. 1b).

The two verification procedures considered by Lidz et al. were labeled the *subtraction procedure* and the *selection procedure*. As we will discuss shortly, these were the two procedures that the design of the task deliberately left available as

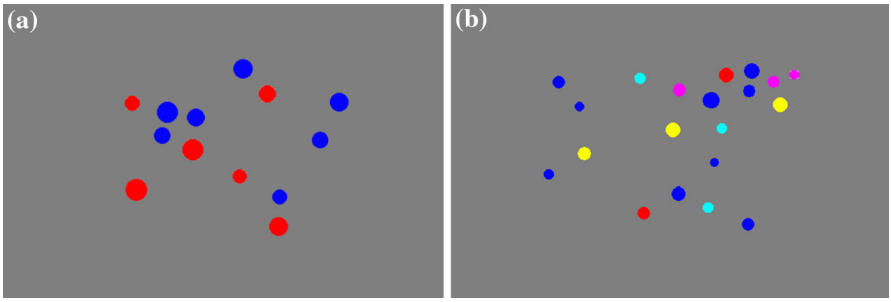


Fig. 1 Sample stimuli from Lidz et al.'s experiment. (Color figure online)

viable options (in the critical trials). The subtraction procedure is described informally in (3).

(3) a. Assess the number of blue dots; let the result be B.
b. Assess the total number of dots; let the result be D.
c. Subtract B from D; let the result be X.
d. If B is greater than X, the final result is True; otherwise, the final result is False.

For example, in a scene with 20 dots of which 14 are blue, this procedure calls for subtracting 14 from 20 to reach the intermediate result 6 (intuitively: the number of non-blue dots), and then comparing 14 with this intermediate result of 6. Since 14 is greater, the final result is True. (Note the obvious but significant point that the procedure in (3) is one of the many that implement the truth condition in (2), and so this final result accords with the fact that the truth condition in (2) is satisfied in the scenario described.)

What Lidz et al. called the selection procedure differs in that there is no role for the total number of dots in this procedure. Instead, it makes use of the individual non-blue color sets. It is shown informally in (4).

(4) a. Assess the number of blue dots; let the result be B.
b. For each non-blue color C_i in the list of colors C_1 through C_n that are present:
Assess the number of dots of this color; let the result be X_i.
c. Sum up $X_1 + X_2 + \ldots + X_n$; let the result be X.
d. If B is greater than X, the final result is True, otherwise, the final result is False.

For example, in a scene with 14 blue dots, 3 yellow dots, and 3 red dots, this procedure calls for summing 3 and 3 to reach the intermediate result 6 (intuitively: the number of non-blue dots), and then comparing 14 with this intermediate result of 6. Since 14 is greater, the final result is True. (Note again that this procedure is one that implements the truth condition in (2).) And to take an important, minimally-different example: in a scene with 14 blue dots and 6 yellow dots, the summing step is trivial and has effectively nothing to do, so running the selection procedure amounts to directly comparing these perceived quantities, again reaching the final result True.

Two formal, non-empirical observations about these procedures are in order. First, they differ only in how they go about computing how many non-blue dots there are: the first and last steps are identical in each case. Second, as the examples above should make clear, the subtraction procedure will run through exactly the same steps in a two-color trial where the blue:yellow:red ratio is 14:6:0 as it will in a three-color trial where this ratio is 14:3:3. Inspecting the internal state of a machine running the subtraction procedure would not provide any information about which of these two kinds of displays had been perceived. In contrast, it is quite a different thing for a machine to run the selection procedure on a three-color 14:3:3 trial than

it is to run the selection procedure on a two-color 14:6:0 trial: among other differences, the former involves a non-trivial summing step whereas the latter does not, and the former draws on three different pieces of external information whereas the latter draws on only two.

With those formal observations noted, there are now three important empirical points to be made about how these two procedures may or may not be used by human subjects in Lidz et al.'s experiment, given the details of the 150 ms display time and the particular kinds of displays used. These follow from generalizations, independently established by previous studies using similar experimental settings, concerning humans' "gut feeling" approximations of numerosity (the Approximate Number System; or ANS; Dehaene 1997; Feigenson et al. 2004; Whalen et al. 1999) and the way these capacities interact with the ability to attend to, or "pick out", a particular subset of the objects in a scene (Wolfe 1998; Treisman and Gormican 1988; Halberda et al. 2006).

First, there is good reason to believe that the human mind is a machine that, while it can run the selection procedure on a two-color 14:6:0 trial given Lidz et al.'s experimental setting, cannot run the selection procedure on a three-color 14:3:3 trial. The reason is that, in the experimental setting used by Lidz et al., participants are able to extract numerosity assessments of at most two proper subsets of the objects on the screen (Halberda et al. 2006). So, slightly more generally, participants are able to use the selection procedure on trials with two colors present, but not on trials with three or more colors present. In the case of the subtraction procedure, on the other hand, this kind of sensitivity to the number of colors present is not possible: any given machine will either be able to run the subtraction procedure on both of these kinds of trials or on neither of them. And as it turns out, there is good reason to believe that the human mind is a machine that *can* run the subtraction procedure on both of these kinds of trials. The reason is that, in addition to the two proper subset approximations mentioned above, participants are able to extract an approximation of the numerosity of the set of *all* dots; so the subtraction procedure is possible on all types of trials (and indeed only makes use of one of the two proper subset approximations that participants can in principle extract).

Second, the previous research indicates that *when* the subtraction procedure and the selection procedure are both viable—i.e. in two-color trials—the selection procedure is more accurate. Recall that the difference between these two procedures is all in how they go about computing how many non-blue dots there are. The reason for the difference in accuracy is that in the case of the selection procedure, the numerosity of the one non-blue color (say, yellow) amounts to the numerosity of the non-blue dots in general (no summing is required), so the numerosity of the non-blue dots is a primitive percept drawn *directly* from the scene. In the case of the subtraction procedure, however, the numerosity of the non-blue dots is computed *indirectly*, via a subtraction operation; consequently, the subtraction procedure uses a noisier approximation of the numerosity of the non-blue dots than the selection procedure does, and is therefore more error-prone. To repeat, however, this discussion of accuracy differences only applies to two-color trials. Table 1 summarizes the consequences, for the efficacy of these two procedures, of varying the number of colors of dots present in a display.

Table 1 The consequences of varying the number of colors present, for the efficacy of the two verification procedures considered by Lidz et al.

	2 colors present	3 colors present	4 colors present	5 colors present
Subtraction procedure	Low accuracy	Low accuracy	Low accuracy	Low accuracy
Selection procedure	High accuracy	Impossible	Impossible	Impossible

Third, and finally, the previous research indicates that, in trials with three or more colors, there is no way to extract a direct, low-noise approximation of the non-blue dots of the sort that the selection procedure extracts in two-color trials. In other words, one might wonder whether, in addition to the two procedures we have been considering, there is a third possibility—one with a step that simply says "Assess the numerosity of the non-blue dots; let the result be X". This is not an option, however, because while it is possible to attend to sets of objects on the basis of color (e.g. all the red dots, all the yellow dots), it is not possible to attend to a set characterized by the negation of a color (e.g. non-blue) or by a disjunction of colors (e.g. red-or-yellow). So in trials with three or more colors, there really is no more accurate alternative than the indirect, and therefore relatively noisy, subtraction procedure.

Lidz et al. measured the proportion of participants' responses that were correct (i.e. in accord with the truth condition in (2)). There are a few candidate hypotheses concerning participants' choices of verification procedure that make straightforward predictions. One possibility is that participants use the subtraction procedure on every trial, in which case we would expect to see no variation in percentage of correct responses as the number of colors present is varied. Another possibility is that participants (try to) use the selection procedure on every trial, which would lead to above-chance accuracy on two-color trials but chance responses on three-, four-, and five-color trials. Finally, one can imagine that participants are able to switch between procedures from trial to trial depending on suitability, in which case we would expect to see the same above-chance accuracy on two-color trials (when the selection procedure is adopted) and lower, but still above-chance accuracy on three-, four-, and five-color trials (when the subtraction procedure is adopted).

The results showed no effect of the varying number of colors present. Participants' responses were equally accurate in two-, three-, four-, and five-color trials. Lidz et al. concluded that participants adopted the subtraction procedure throughout.[4] In particular, participants used the subtraction procedure even in two-color trials, where the more accurate selection procedure was a viable alternative.

The main point of Lidz et al.'s paper is that this would be a surprising result if one took the endpoint of a full and complete understanding of a sentence to be a truth condition. On such a view one would expect the choice of verification procedure, among those viable given task constraints, to be constrained only by the requirement that the chosen procedure must implement that truth condition. This would leave participants' neglect of the selection procedure in two-color trials

[4] Furthermore, the results are as predicted by this hypothesis even in the finer-grained details: based on independently-proposed models of the ANS, this hypothesis also predicted the particular rate at which participants' proportion of errors increased as the ratio of blue to non-blue dots approached 1:1.

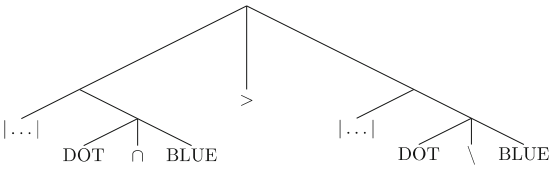
mysterious. Lidz et al. hypothesized that the source of this neglect of the selection procedure—recall from the reasoning above that the source was not task constraints—was the meaning of the target sentence. For the meaning of the target sentence to have such an effect, its properties cannot be exhaustively characterized by a truth condition; rather, sentence meanings must be more finely individuated than truth conditions are.[5]

To begin to flesh out this idea, Lidz et al. suggest that we take sentence meanings to be structured objects, such that (5a) and (5b) are distinct, though truth-conditionally equivalent, candidate meanings that theorists can argue for and against; and they propose that (all else being equal) speakers are biased towards verification procedures that directly compute the operations represented in these structured objects (they dub this the "Interface Transparency Thesis"). Note that no operations are represented in any truth condition.

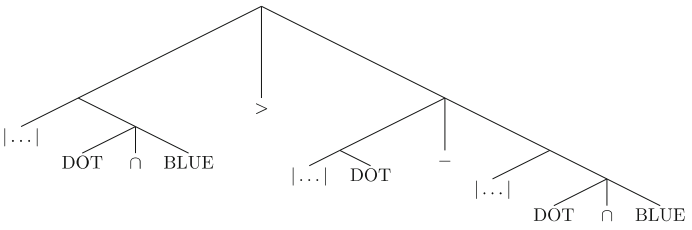
(5) a. | DOT ∩ BLUE | > | DOT \ BLUE |
b. | DOT ∩ BLUE | > | DOT | – | DOT ∩ BLUE |

To bring out more clearly the distinctions between the two objects in (5), we might write them instead as in (6).

(6) a.



b.



The guiding idea is that there is a part of (6b) that transparently corresponds to the subtraction operation which participants surprisingly persisted with in the experiment, while there is no such subtraction operation on numerosities/cardinalities in (6a). So, by hypothesizing that (6b) and not (6a) is the relevant mental object, in combination with the ITT, we can formulate a candidate explanation for the results of Lidz et al.'s experiment. In effect, a sentence meaning, in addition to

[5] More correctly speaking, the conclusion that this line of reasoning leads to is that sentence meanings are individuated *not as coarsely as* truth conditions. The important point which we are focusing on here is that two sentences might have different meanings and yet share the same truth condition, and so the relation between sentence meanings and truth conditions is at least many-to-one; but in general one might also want to consider the possibility that a sentence meaning does not uniquely determine a truth condition (the truth condition instead being the result of an interaction between the sentence meaning and various contextual factors), in which case the relation would be many-to-many.

determining a certain function from contexts to truth values (though see footnote 5), has some "verificational weight" which, when combined with all the other factors bearing on participants' choice of verification procedure in Lidz et al.'s experiment, somehow tips the scales in favor of the subtraction procedure in two-color trials, outweighing the fact that this procedure's accuracy is relatively poor. See Lidz et al. (2011) and Pietroski et al. (2011) for further discussion.

This hypothesis leaves unanswered many questions about what scale-tipping effect this verificational weight might have in other settings. In particular, there might be other settings where it does not outweigh other, more obvious factors that contribute to choices of verification procedure. (A moment's reflection suggests that this is almost certainly the case.) There are also many details to be worked out concerning the relationship between the structured expressions in (6) and the verification procedures mentioned earlier in (3) and (4). It may also be that the ITT-based approach that Lidz et al. suggest turns out to be an unfruitful way to theorize about the factors that affect speakers' choice of verification procedures. Be that as it may, in order to understand the relationship between Lidz et al.'s study and KSH, the important point to take in is simply that Lidz et al. conducted an experiment to detect which of two truth-conditionally equivalent verification procedures participants used in responding to the sentence in (1), and then proposed a possible explanation for the pattern of choices that they found.

3 KSH'S investigation of truth conditions

KSH's main claim is that the sentence in (1) is ambiguous. In addition to the proportional truth condition discussed in Sect. 2, KSH argue that (1) has what they call the superlative truth condition, which requires that there be more blue dots than there are dots of any other color. The relationship between these two readings is analogous to the relationship between the most salient reading of (7a), which expresses a proportional-type truth condition, and the most salient reading of (7b), which expresses a superlative-type truth condition.

(7) a. John painted most of the dots.
    b. John painted the most dots.

For KSH, part of the significance of this additional reading stems from the fact that it provides evidence for the hypothesis that both 'most' in (7a) and 'the most' in (7b) are built from the same ingredients, namely a gradable predicate 'many'/'much' and the superlative operator '-est', but are structured differently and therefore interpreted differently (Hackl 2009). This hypothesis predicts that the string in (1) should have two analyses, one of which gives rise to the proportional reading and the other of which gives rise to the superlative. By contrast, the superlative reading of (1) would be unexpected on the view that 'most' and 'the most' in (7) are simply unrelated lexical primitives.

For our purposes here, however, we can leave aside the debate between the lexical and structural views of 'most'. This is not to deny that the presence of the

superlative reading provides support for the structural view, but simply because we want to focus on the claim that the superlative reading is present (whatever the theoretical ramifications of this claim) and on the experimental evidence for this claim, since this is where the apparent conflict with Lidz et al.'s earlier work lies.[6]

The task used by KSH is another picture verification task: presented with a picture showing an array of dots, participants were asked to judge a target sentence true or false by pressing buttons on a keyboard. The target sentence was (1) for participants in one condition, and was 'More than half of the dots are yellow' in the other; we will focus on the former. The task differed from Lidz et al.'s, however, in that the picture was displayed without any time limit, so participants could look at the array of dots for as long as they wanted before responding.

Like Lidz et al., KSH varied the ratio of blue to non-blue dots and the number of colors present: on each display there were either two colors (blue and yellow) or three colors (blue, yellow, and red) present. In addition, KSH manipulated the makeup of the non-blue dots. Among three-color displays where the blue:non-blue ratio was 9:11, for example, there were three different ways in which the non-blue dots could be divided between yellow and red: the blue:yellow:red ratio was either 9:6:5 ("balanced"), 9:8:3 ("mildly balanced"), or 9:10:1 ("unbalanced"). In all these example cases the 9:11 ratio makes the proportional reading false, but the truth value under the superlative reading varies: this reading is true in 9:6:5 and 9:8:3 displays, but false in 9:10:1 displays.[7]

KSH point to two patterns in the experimental results as evidence for the superlative truth condition. First, pictures that satisfied the superlative truth condition were judged true more often than other pictures with the same blue:non-blue ratio that did not satisfy this truth condition. In reference to the examples above, this amounts to higher rates of 'true' responses in the 9:6:5 and 9:8:3 displays than in the 9:10:1 displays—and in the two-color 9:11:0 displays—despite the fact that the proportional reading is false in all three cases by virtue of the common 9:11 blue:non-blue ratio. Furthermore, the rate of 'true' responses increased with the ratio of blue to yellow (i.e., the most numerous non-blue color) dots. In other words, there were more 'true' responses for 9:6:5 displays than for 9:8:3 displays.

Despite these trends, the rates of 'true' responses to displays that satisfied the superlative truth condition were "overall rather low and, in fact, almost without exception below 50%" (p. 137). This would be unexpected if the two readings were equally salient for all speakers, but KSH offer the explanation that (a) the superlative reading is "latent and often masked by the more dominant proportional reading" (p. 137), and (b) the superlative reading is only available to a certain portion of participants (56 of 135 participants, according to a classification that KSH use). The argument for the ambiguity of 'most' is therefore somewhat indirect, but let us grant

[6] KSH suggest that Lidz et al. present a version of the lexical view of the relation between (7a) and (7b), but this was not intended. While Lidz et al. do not mention the possibility of decomposing 'most', this is because their questions concerned whole-sentence meanings, rather than how these meanings are composed.

[7] In Lidz et al.'s experiment, the non-blue dots were always evenly distributed among the other colors, corresponding to KSH's "balanced" condition.

KSH's point that if some participants were adopting the superlative truth condition some of the time, then this would indeed explain the two trends identified above.

In addition to making this argument for the superlative truth condition, however, KSH go on to claim that a comparison of their findings with those of Lidz et al. "casts doubt on the generalizability of the results [of Lidz et al.] and on the robustness of the conclusions drawn from that study" (p. 151). We disagree with this further point. In the next section we will explain why, and describe how we see the two studies' findings to be compatible.

4 Synthesizing the findings

KSH suggest that the results of the two experiments can be reconciled, but only at the expense of the conclusions that Lidz et al. drew from their results. So, to be clear, KSH's claim is not that there is no internally consistent theory that will account for the two sets of results, but rather that the conclusions we should draw from the two taken together conflict with the conclusions that Lidz et al. drew from theirs in isolation. We take issue not with the claim that a unified explanation is possible, but rather with the claim that this unified explanation conflicts with Lidz et al.'s conclusions.

There are two points of apparent conflict between the studies. The first is the fact that Lidz et al. operated under the assumption that the target sentence has only the proportional truth condition. This assumption obviously conflicts with KSH's main claims, but it was not part of Lidz et al.'s intended contribution to provide evidence that the target sentence was unambiguous. The assumption turned out to be harmless because, as KSH observe, participants in Lidz et al.'s experiment did end up uniformly adopting the proportional truth condition. KSH's explanation for this is the fact that the superlative reading is "latent and often masked by the more dominant proportional reading" (p. 137). What this would mean is that the question of why participants ended up uniformly using the proportional truth condition rather than the superlative one is perhaps somewhat murkier than Lidz et al. assumed. But the other question, namely why they uniformly adopted the subtraction procedure as their method of computing this proportional truth condition—which was Lidz et al.'s concern—is not affected by this murkiness of the first question. So KSH's discovery of the superlative truth condition is neither in conflict with Lidz et al.'s conclusions about verification procedures nor disruptive of the logic used to reach those conclusions.[8]

The second and more subtle point of apparent conflict between the two studies concerns the fact that KSH found that rates of 'true' responses were affected by the number-of-colors manipulation. Although Lidz et al. found no significant differences between two-, three-, four-, and five-color trials, KSH found higher rates of 'true' responses in (for example) three-color displays where the blue:yellow:red ratio was 9:6:5 than in two-color displays where this ratio was

[8] Lidz et al. is not disputed by the finding that 'most' is two-ways ambiguous any more than KSH's claim would be significantly disputed by a future finding that 'most' is in fact three-ways ambiguous, with some new third reading in addition to the two that KSH identify.

9:11:0 (along with intermediate rates of 'true' responses at the intermediate ratios of 9:8:3 and 9:10:1, which do not correspond to cases that Lidz et al. tested). This does not itself constitute any conflict or inconsistency, since the two tasks were different, as KSH point out. So one might expect that this should be the end of the story, with each of the two studies' conclusions left to stand on their own: KSH's conclusions about the target sentence's range of truth conditions, and Lidz et al.'s conclusions about the verification procedures speakers use with the proportional truth condition.

In discussing the effects of the differences between the experimental tasks, however, KSH run together the two apparent conflicts just mentioned. They therefore end up conflating the fact that participants in Lidz et al.'s experiment did not adopt the superlative truth condition with the fact that the participants did not use the selection procedure. KSH "suggest that [the design of Lidz et al.'s experiment] and the specific task demands that came with it biased participants toward using a verification strategy such as [the subtraction procedure]" (p. 151). One way to understand this suggestion is as an explanation for the first apparent conflict mentioned above, concerning ambiguity: if "a verification strategy such as the subtraction procedure" simply means "a verification strategy that implements the proportional truth condition", then this reiterates the point made above that, for whatever reason, the Lidz et al. task prompted participants to use the proportional truth condition rather than the superlative one. Understood this way, we have no issue with the suggestion.

The only way to understand KSH's suggestion in a way that would cast doubt on the conclusions drawn by Lidz et al., however, is to take it to imply that the design of the experiment biased participants towards using the subtraction procedure rather than the selection procedure on the critical two-color trials. We see no justification for suspecting this, and plenty of reasons to doubt it. First, there is no reason to believe that the properties of the individual trials created any such bias: as mentioned above, and as discussed at length in Lidz et al., the individual trials (display time, dot layout, etc.) were specifically designed to ensure that the two procedures under consideration were both feasible (on the critical two-color trials), and furthermore that the selection procedure was the more accurate. Second, KSH raise the possibility that having a single target sentence throughout the experiment (rather than including trials with a variety of different target sentences) might lead participants to choose a single verification procedure for the experiment as a whole rather than choosing procedures on a trial-by-trial basis. Specifically, the fact that *not all* trial types were compatible with the selection procedure might lead participants to put it aside altogether and use it on *none* of the trials. Lidz et al. note (footnote 7), however, that a comparison with a very similar previous experiment (Pietroski et al. 2009) indicates that participants use the subtraction procedure even when only two-color trials are present.[9] So the effect cannot be attributed to participants "sticking with" a verification procedure that will work for all trial

[9] Pietroski et al. (2009) also provides evidence that in this kind of experiment with an invariant target sentence, participants can adopt different verification procedures for certain suitable trial types, namely the "sorted columns" trials on that experiment.

types. We maintain, then, that participants in Lidz et al.'s experiment consistently used the subtraction procedure for some reason that can not be attributed to the particulars of the individual trials nor to the overall single-target-sentence nature of the experiment.

The crucial point for addressing the apparent conflict over the number-of-colors manipulation is that one would expect an effect of this manipulation if *either* (i) participants were adopting the superlative truth condition, *or* (ii) participants were adopting the proportional truth condition and using the selection procedure to evaluate it. Lidz et al. found no such effect in their experiment, and concluded that participants were adopting the proportional truth condition and using the subtraction procedure to evaluate it. Lidz et al. took one half of this conclusion to be interesting, and took the other half to not be. The interesting part was the fact that participants were using the subtraction procedure (rather than (ii) above): this was interesting because there was an alternative procedure, the selection procedure, that was viable given task constraints, truth-conditionally equivalent and more accurate, and yet not used. The part Lidz et al. took to be uninteresting was the fact that participants were adopting the proportional truth condition (rather than (i) above): they had no particular alternative to this in mind, and were not investigating any questions concerning the target sentence's range of truth conditions. KSH take their results to indicate that this choice of truth condition was not a foregone conclusion in the way that Lidz et al. assumed, and was perhaps attributable to the experimental setup. But this does not mean that the interesting finding, the neglect of the selection procedure in favor of the noisier subtraction procedure, was attributable to the experimental setup.

KSH do not appear to carefully consider the distinction between questions about truth conditions and questions about verification procedures. They write, for example, that "if 'most' is unambiguously a proportional determiner and truth-conditionally equivalent to 'more than half'…we again expect any experimental manipulation to affect these two determiners equally" (p. 151). This neglects possibilities of exactly the sort that Lidz et al. argued for: for example, the possibility that 'most' might indeed be truth-conditionally equivalent to 'more than half', and yet come with some sort of "verificational weight" (not shared by 'more than half') which sometimes tips the scales in favor of certain verification procedures that one might not otherwise expect to be used. Neglecting this possibility would leave one unable to appreciate the distinction between the number-of-colors manipulation as a probe of whether the superlative (rather than proportional) truth condition is being adopted, and as a probe of whether the selection (rather than subtraction) procedure is being used to evaluate the proportional truth condition.

To clarify the point it is perhaps worth noting that although KSH speak of "verification strategies" and "verification profiles", they are not trying to make any claims about, or considering any candidate hypotheses about, the verification procedures used by participants in their experiment (except in the trivial sense that any truth condition of course picks out the range of imaginable verification procedures that implement it). This is made quite explicit: KSH (p. 124) say that they "use the term 'proportional verification strategy' to refer to the idea that

speakers verify a sentence according to a proportional truth condition" (and likewise for 'superlative verification strategy'). Their considerations of different verification strategies remain at this level of granularity, without considering finer-grained subclasses of verification strategies, such as the distinction between selection and subtraction procedures. This is entirely appropriate and understandable given their task, because it would be difficult to enumerate the large range of verification procedures that are viable options in the unlimited display time; and it is also entirely sufficient for their goals, which concern claims about truth conditions. But the fact that the two papers used terms like "verification strategy" and "verification procedure" in these distinct ways perhaps masks the differences in their approach and emphasizes the appearance of conflict between them. Although both experiments comprised what might be called a "verification task", and used the accuracy of responses as the dependent variable, KSH compare the observed accuracies with what would be predicted if various truth conditions were being adopted, whereas Lidz et al. compare the observed accuracies with what would be predicted if various verification procedures were being adopted. Note also that if Lidz et al.'s findings are taken seriously (and we still see no reason why they should not be) then both of these kinds of questions are relevant to theories of linguistic meaning, for the reasons mentioned in Sect. 2 above.

In light of this discussion it may be useful to consider the role of the main difference between the two tasks, the strict 150 ms display time that was imposed on participants in the Lidz et al. experiment. It was this restricted display time that allowed Lidz et al. to enumerate a restricted range of candidate verification procedures that were cognitively feasible in that experimental setting, and identify the predicted accuracy profile of each one. If one is investigating truth conditions, as KSH were, then there is no particular need to restrict participants to some small, tractable handful of verification procedures; indeed, if one is investigating truth conditions, imposing such a short display time would perhaps be an odd thing to do, since it would probably lower participants' accuracy and add noise to the data. But for the questions Lidz et al. were investigating, the 150 ms display time was a crucial ingredient in the logic used to interpret the results of the experiment, not an aspect of the design that got in the way.[10]

5 A note on the superlative truth condition

For the purposes of distinguishing the goals of the two studies from each other, we granted KSH the conclusion that (1) is ambiguous for some speakers and that this conclusion is warranted given the way some of their participants' responses were affected by the number-of-colors manipulation. However, now that the logic of the

[10] We agree with KSH's statement (pp. 152–153) that data of the sort discussed here must be interpreted by "combining hypotheses about the semantics of the studied expression with a (ideally independently justified) theory of the task", but disagree with the implication that Lidz et al. did not do so. The 150 ms display time was imposed precisely in order to allow the results to be interpreted with reference to an independently justified theory of the task.

two studies has been clarified, we would note that this argument for the superlative truth condition becomes less convincing when one considers other kinds of evidence that in principle might have been used to construct more direct arguments for the same conclusion.

The sentences in (8), for example, can be understood in ways that don't make the speaker's position contradictory due to the (lexical or structural) ambiguity of the bracketed embedded clauses. For any speakers for whom (1) is similarly ambiguous between the proportional and superlative truth conditions – as KSH argue that it was for around 40% of their participants – one might expect the sentence in (9) to have the same kind of non-contradictory reading.

(8) a. I affirm that [my car is a lemon], but I deny that [my car is a lemon].
b. I affirm that [I went to the bank], but I deny that [I went to the bank].
c. I affirm that [I saw a man with a telescope], but I deny that [I saw a man with a telescope].

(9) I affirm that [most of the dots are blue], but I deny that [most of the dots are blue].

If necessary, (9) could be accompanied by a context that favors the superlative reading (although note that this does not appear to be necessary for the examples in (8)). For example, we could imagine a competition where a blue team, a red team, and a yellow team are attempting to paint as many dots as possible with their respective color, and a team can win a prize by painting more dots than any other team. We would then expect that, for speakers with a superlative reading, (9) could be uttered, without contradiction, by a judge explaining the choice to award the prize to the blue team. (Since this scenario provides no contextual support for the proportional truth condition, the second clause might have a pragmatically odd "out of the blue" character, but no contradiction is expected.)

Whatever the empirical status of (9) turns out to be, a complete account of the semantics of *most* will of course need to account for both that data and KSH's experimental findings. And recall that those experimental findings neither conform perfectly to the predictions of the proportional-only hypothesis (as KSH stressed), nor conform perfectly to those of a straightforward version of the ambiguity hypothesis according to which both readings are equally "accessible". This is understandable given the novelty of the task, since there are plenty of unknowns mediating the relationship between participants' responses and the hypotheses being tested which may play a role in explaining the results. For the ambiguity hypothesis, some missing puzzle pieces need to be added in order to explain the fact that even those speakers for whom the target sentence was apparently ambiguous used the superlative truth condition only some of the time; Kotek et al. (2011, p. 124) suggest that the "processing difficulty associated with a particular reading" may contribute to when this reading is and is not accessed, but without independent measures of such processing difficulty this cannot provide much explanatory force. For the proportional-only hypothesis, the missing pieces must provide some explanation for

the increase in 'true' responses in the balanced conditions of KSH's experiment; one form this may take would be positing a verification procedure implementing the proportional truth condition whose error profile yields the observed patterns—but again, without spelling out further details this cannot provide much explanatory force.[11]

We are wary of rejecting the proportional-only hypothesis on the basis of the currently available evidence.

6 Conclusion

KSH's evidence for a superlative truth condition for 'most' has no significant bearing on the conclusions reached by Lidz et al. The independence of the two studies' claims stems from the fact that Lidz et al. were investigating speakers' choices among alternative truth-conditionally equivalent verification procedures, whereas KSH were (despite similarities in terminology) investigating coarser-grained questions concerning the range of truth conditions that 'most'-sentences have. Although both studies manipulated the number of colors of dots in their experimental stimuli, this manipulation played a different role in each case. For Lidz et al., the absence of an effect of this manipulation in their experimental setting was taken as evidence that participants opted for the subtraction procedure rather than the selection procedure; for KSH, the presence of an effect in their experimental setting was taken as evidence for the superlative truth condition.

References

Dehaene, Stanislas. 1997. *The number sense: How the mind creates mathematics*. New York: Oxford University Press.

Feigenson, Lisa, Stanislas Dehaene, and Elizabeth Spelke. 2004. Core systems of number. *Trends in Cognitive Science* 8: 307–314.

Gallistel, C.R., and Adam Philip King. 2009. *Memory and the computational brain*. Malden, MA: Wiley-Blackwell.

Hackl, Martin. 2009. On the grammar and processing of proportional quantifiers: *most* versus *more than half*. *Natural Language Semantics* 17: 63–98.

Halberda, Justin, Sean F. Sires, and Lisa Feigenson. 2006. Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science* 17: 572–576.

Kotek, Hadas, Yasatada Sudo, and Martin Hackl. 2015. Experimental investigations of ambiguity: the case of *most*. *Natural Language Semantics* 23: 119–156.

Lidz, Jeff, Paul Pietroski, Tim Hunter, and Justin Halberda. 2011. Interface transparency and psychosemantics of *most*. *Natural Language Semantics* 19: 227–256.

Marr, David. 1982. *Vision*. New York: W.H. Freeman and Company.

Pietroski, Paul, Jeff Lidz, Tim Hunter, and Justin Halberda. 2009. The meaning of *most*: Semantics, numerosity and psychology. *Mind & Language* 24(5): 554–585.

Pietroski, Paul, Jeff Lidz, Justin Halberda, Tim Hunter, and Darko Odic. 2011. Seeing what you mean, mostly. In *Syntax and semantics 37: Experiments at the interfaces*, ed. Jeff Runner, 187–224. New York: Academic Press.

[11] Note that nothing in Lidz et al.'s proposal makes any concrete predictions about the choice of verification procedure that participants will make in KSH's experimental setup, even in the context of the assumption that (1) has only the proportional truth condition.

Treisman, Anne, and Stephen Gormican. 1988. Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review* 95(1): 15–48.

Whalen, John, C.R. Gallistel, and Rochel Gelman. 1999. Non-verbal counting in humans: The psychophysics of number representation. *Psychological Science* 10: 130–137.

Wolfe, J.M. 1998. Visual search. In *Attention*, ed. H. Pashler, 13–73. London: University College London Press.