In: Proceedings of the 15th Texas Linguistic Society, edited by Christopher Brown, Qianping Gu, Cornelia Loos, Jason Mielens, and Grace Neveu, pp. 39-52.
URL: http://tls.ling.utexas.edu/2014tls/TLS15-Proceedings.pdf

Truth value judgments vs. validity judgments*

Elizabeth Coppock
University of Gothenburg and Swedish Collegium for Advanced Study
eecoppock@gmail.com

Abstract

This paper undertakes a direct comparison between two methodologies for getting at semantic intuitions: (i) validity judgments, where subjects judge the validity of arguments, e.g. *There are three bananas; therefore there are at least three bananas,* and (ii) picture verification tasks (also known as 'truth judgment tasks'), in which, for example, one sees a picture of three bananas and judges a statement like *There are at least three bananas*. It has been suggested that validity judgment tasks are more sensitive to ignorance implicatures than picture verification tasks, but these two methods have not been compared directly using comparable stimuli. The present work aims to close that gap. The results show that validity judgment tasks do not in fact robustly pick up on ignorance implicatures, so they cannot be relied upon for that, although both validity judgment tasks and truth value judgment tasks are sensitive to violations of particularly strong pragmatic requirements. In general, the two kinds of tasks gave quite similar results. This raises the question of why validity judgment tasks sometimes pick up on ignorance implicatures and sometimes do not.

1 Introduction

Recent work has suggested that *validity judgment tasks*, where subjects evaluate the validity of arguments, differ from *picture verification tasks* (or *truth judgment tasks*), where subjects judge sentences as true or false vis-à-vis a depicted scenario (Coppock & Brochhagen 2013a; henceforth C&B). C&B suggest that "truth-value judgments are less sensitive to pragmatic infelicity than inference judgments," and in particular, "ignorance implicatures do not affect truth-value judgments even though they do affect inference judgment tasks," although some pragmatic principles are so strong that violations of them "can cause true sentences to be judged as false." These conjectures were not based on a sufficiently controlled comparison between the two types of task, however. The present paper aims to provide one, so as to gain a better understanding of what these tests can be used to diagnose.

The story begins with the validity judgment experiments carried out by Geurts et al. (2010), in which participants were asked to judge the validity of inferences from one sentence to another. The results showed a difference between comparative modifiers (*more than* n, *less than* n) and their mathematically equivalent superlative counterparts (*at least* $n+1$, *at most* $n-1$). An argument like the following was unanimously judged to be a valid argument:

*Thanks to Alexander Coppock for help with plotting the data in R. This research was made possible through funding from Riksbankens Jubileumsfonds *Pro Futura Scientia* program, hosted by the Swedish Collegium for Advanced Study.

(1) Berta had three beers. Therefore, Berta had more than two beers.

In contrast, only 50% of their participants judged the following to be a valid argument:

(2) Berta had three beers. Therefore, Berta had at least three beers.

A similar constrast was found between *fewer than* and *at most*. While it was not unanimous, the overwhelming majority of participants considered the following a valid argument (93%):

(3) Berta had three beers. Therefore, Berta had fewer than four beers.

But only 61% considered the variant with *at most* valid:

(4) Berta had three beers. Therefore, Berta had at most three beers.

These findings can be understood broadly under the uncontroversial assumption that superlative modifiers (*at least*, *at most*) carry some kind of ignorance implication that *more* and *less* do not. To say *Berta had at least three beers* signals at some level that one does not know how many beers Berta had. To say *Berta had more than two beers* does not carry the same sort of ignorance implication.

There are several classes of theories about this ignorance implication. Geurts & Nouwen (2007) proposed that it was a logical entailment. On Geurts & Nouwen's (2007) ignorance-as-entailment view, *Berta had at least 3 beers* is true if and only if:

The speaker considers it necessary that Berta had 3 beers or more
and considers it possible that Berta had more than 3 beers.

Although details differ, the majority of views on superlative modifiers take the ignorance component to be an implicature (Büring 2008, Cummins & Katsos 2010, Cohen & Krifka 2011, Biezma 2013, Coppock & Brochhagen 2013b, Mayr 2013, Schwarz 2013, to appear: i.a.). On an ignorance-as-implicature view, *Berta had at least 3 beers* is true if and only if

Berta had 3 beers or more.

The ignorance implication is not part of the truth conditions on this view.

One instantiation of the ignorance-as-implicature view is given by Büring (2008) (followed by Cummins & Katsos 2010 and Biezma 2013). On this view, *at least p* 'amounts to a disjunction' between *p* and *more than p*, and there is an 'implicature schema' that says, if a speaker says *A or B*, then the speaker considers both *A* and *B* to be possible. As Coppock & Brochhagen (2013b) discuss, this raises the question of what it means for a speaker to 'say' *A or B*. Clearly there is no *or* in the surface string in a sentence containing *at least*. Büring describes no syntactic transformation that would transform the surface string to to an LF containing an LF. The only place where an 'or' shows up in Büring's theory is in the characterization of the denonation of *at least*, where it is part of the meta-language, one cannot hang implicatures on distinctions made in the meta-language.

Coppock & Brochhagen (2013b) offer a view building on a similar intuition which is not subject to this problem. On the Coppock & Brochhagen 2013b view, saying *at least p* is not saying *p or more than p*, but superlative modifiers have an important property in

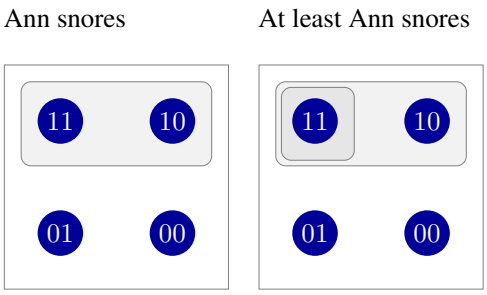


Figure 1: *Ann snores* vs. *At least Ann snores*

common with disjunctions, one which is also shared by questions: They raise issues. This issue-raising property is expressed with the use of inquisitive semantics (Groenendijk & Roelofsen 2009).

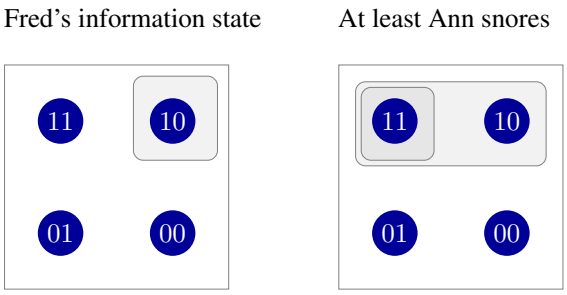
To illustrate, let us consider a simple example, with four worlds: Ann and Bill snore (w_{11}), Ann snores and Bill doesn't (w_{10}), Ann doesn't snore and Bill does (w_{01}), and neither snore (w_{00}). The sentence *Ann snores* picks out the possibility that Ann snores, which is true in two worlds: the one where both Ann and Bill snore, and the one where Ann snores and Bill does not. The sentence *At least Ann snores* (with focus on *Ann*) picks out two possibilities: the possibility that Ann snores, and the possibility that *something stronger* holds – in this case, that Ann and Bill snore. This is depicted in Figure 1. So the two sentences are true in the same set of possible worlds, but the *at least* sentence differentiates among these possible worlds more finely.

This theory makes it possible to articulate the ignorance-as-implicature view as follows. Let us say that a sentence meaning is *interactive* if it raises an issue, and let us define this technically by saying that ϕ is interactive iff the denotation of ϕ contains more than one possibility. An ignorance implicature arises with an interactive (or issue-raising) sentence based on the following reasoning: If the speaker already knew how to resolve the issue, then there would be no reason to bring it up. An issue raised by a speaker should be unresolved in the speaker's mind. (In fact this is a bit strong, because there may be many reasons to raise an issue; cf. Sven Lauer's discussion of 'Need A Reason' implicatures, but we simplify here for the sake of discussion.) Call the set of possible worlds that are epistemically accessible to the speaker the speaker's *information set*, and let us say that a sentence is *interactive in an information set* if, when we restrict its denotation to the worlds in the information set, the result still contains multiple possibilities (Groenendijk & Roelofsen 2009). The Gricean maxim by which the implicature arises can then be codified as follows (Coppock & Brochhagen 2013b):

(5) Maxim of Interactive Sincerity
If ϕ is interactive, then ϕ is interactive in the speaker's information set.

This can be glossed, 'don't bring up an issue that you already know how to resolve'. For example, if Fred's information state is as in Figure 2, then he should not assert *At least Ann snores* (with focus on *Ann*), because the issue is already resolved in his information set.

Another class of theories of superlative modifiers that is not subject to Coppock &



Fred should not assert *At least Ann snores*.

Figure 2: A violation of the Maxim of Interactive Sincerity

Brochhagen's (2013b) criticism of Büring's analysis appeals to lexically specified alternative expressions of the kind that figure in Horn scales ('scalar alternatives', also sometimes called 'formal alternatives'). This group includes both neo-Gricean theories, on which the relevant implicatures are computed on the basis of Gricean reasoning about alternative expressions (incl. Schwarz & Shimoyama 2011, Schwarz 2013, and Kennedy 2015) and theories based on the grammatical view of scalar (and ignorance) implicatures (Chierchia et al. 2008), on which they owe their presence to a silent exhaustivity operator, advocated by Mayr (2013). Many such theories assume the relevant alternatives for *at least* n to be n and *more than* n (Schwarz & Shimoyama 2011, Schwarz 2013, Kennedy 2015), thereby implementing Büring's idea that *at least* n sentences 'amount to a disjunction' between n and *more than* n. Mayr (2013) assumes that the alternatives are *at least* n, *at least* $n+1$, *at least* $n+2$, etc., along with *at most* n, *at most* $n+1$, *at most* $n+2$, etc. There is ongoing discussion under this umbrella as to what alternatives should be stipulated. Common to these approaches however is that ignorance implicatures are derived on the basis of a so-called 'symmetry problem', which prevents a hearer from concluding that a speaker disbelieves any of the alternatives. As far as I know, there has not been much work done comparing the predictions of these various approaches, although McNabb & Penka (2014) report a set of experimental findings that partially go against all of the theories on the market. It thus appears to remain an open question which of these approaches is to be preferred, if any.

But the purpose of this paper is not to adjudicate among theories of superlative modifiers. This paper is focussed on methodology. The goal is to better understand the nature of the validity judgment task used by Geurts et al. (2010), and compare it more directly to *picture verification tasks*, where truth-judgments are given with respect to a depicted scenario. These appear to give slightly different results, as described in the next section, and this paper aims to make a small step towards understanding why.

2 Picture verification tasks

Recall Geurts et al.'s (2010) results:

(6) a. *Liz had 3 beers* $\Rightarrow$ *Liz had more than 2 beers.* 100%

b. *Liz had 3 beers* $\Rightarrow$ *Liz had at least 3 beers.* 50%

c. *Liz had 3 beers ⇒ Liz had fewer than 4 beers* 93%

d. *Liz had 3 beers ⇒ Liz had at most 3 beers* 61%

Prima facie, such results may be taken to support the view of ignorance as entailment. Indeed, this is how Geurts et al. (2010) argue. But this interpretation of the results depends on a particular view of validity and how validity judgments work: That validity amounts to *truth-preservation* ('A therefore B' is true if and only if B is true whenever A is true), and that validity judgments reflect this view of validity.

Kaplan (1999) promotes a view of validity as *information delimitation*. According to his intuitions, the inference in (7) is valid, but the inference in (8) is not.

(7) That damn Kaplan was promoted.
Therefore, Kaplan was promoted. (valid)

(8) Kaplan was promoted.
Therefore, that damn Kaplan was promoted. (valid?)

The premise is true under all the same circumstances as the conclusion in both cases. But the conclusion contains more information in the case of (8), namely information about the speaker's attitude toward Kaplan. This can explain a feeling of resistance to saying that (8) is a valid inference.

The view of validity as information delimination is one that can be combined with the ignorance-as-implicature view to account for the pattern of validity judgments found by Geurts et al. (2010). Sentences with superlative modifiers carry extra information in the form of ignorance implicatures, so when they appear as the conclusion of an argument and this ignorance is not present in the premise, there is a hesitancy to judge the argument as valid.

Some authors have, in effect, expressed the intuition that the arguments in (6) are truth-preserving. Regarding Geurts & Nouwen's (2007) proposal that superlative modifiers semantically encode speaker ignorance, Cohen & Krifka (2011) write:

> Suppose John committed exactly four traffic violations, but nobody knows this, not even the police (who are the authority on the subject), and not even John himself. Then, it would still be truth that he committed at least three traffic violations, and these truth values depend only on what actually happened, not on anybody's beliefs.

If we *suppose* that John committed four traffic violations, and then ask ourselves whether it is true or false that John committed at least three (or four) traffic violations, then our intuitions become a bit clearer than they were in the case of validity judgments.

This passage suggests an alternative way of getting at speakers' intuitions about information-preservation: rather than asking for validity judgments, *depict* a scenario where the premise is true, and then ask whether the conclusion is true or false.[1] The experiments

[1] In fact, Geurts et al. (2010) did carry out this kind of experiment. They used sentences of the form 'There are Q N Xs', where Q was *exactly*, *at least*, *at most*, *more than* or *fewer than*; N was a number, and X was a letter (either A or B). The sentences were accompanied by a display consisting of some number of instances of the relevant letter, either A or B, and the participants were asked to evaluate the truth of the sentence. They do not report 'accuracy' results for this experiment, however, only response time.

There are at least 6 dogs in the picture.

True
False

There are at most 3 bananas in the picture.

True
False

Figure 3: Sample stimuli used in Coppock & Brochhagen (2013a)

reported in C&B work in this way. Some sample stimuli are shown in Figure 3. A picture with n objects (six dogs, four bananas, etc.) is shown along with a sentence and the participant is asked to judge whether the sentence is true or false. The sentences were all of the form *There are ____ [nouns] in the picture*, where the blank is filled in with a modified numeral.

In their first experiment, C&B presented subjects with pictures containing three, four, five or six objects of a given type (e.g., six puppies), in one of eight conditions, given n objects in the picture:

1. *at most* n
2. *fewer than* $n+1$
3. *at least* n
4. *more than* $n-1$
5. *at most* $n-1$
6. *fewer than* n
7. *at least* $n+1$
8. *more than* n

Conditions 1-4 are ones where the sentence is 'mathematically true' as it were, given the depicted scenario, i.e., true on an interpretation of *at least 3* as ≥ 3, etc. Conditions 5-8 are ones where the sentence is 'mathematically false'. Since subjects consistently rated the sentences in the 'mathematically false' condition as false, they can be ignored.

In this experiment, C&B found that their participants' truth judgments perfectly matched the 'mathematical' predictions: They were at ceiling (with sentences very nearly unanimously judged 'true') in the 'mathematically true' conditions, and at floor (nearly unanimously 'false') in the 'mathematically false' conditions. These results accord with Cohen and Krifka's intuition that the *at most* n sentence is true in a situation where there are n objects of the relevant type, regardless of who believes or knows what. In other words, it supports the idea that the inference is in fact truth-preserving.

So C&B's picture verification task results do not match Geurts et al.'s inference judgment task results (summarized above in (6)). What can explain this contrast? One imaginable hypothesis is that picture verification tasks are just completely impervious to impli-

catures. C&B's second experiment shows this not to be the case; picture verification tasks do pick up on some implicatures. In their second experiment, the stimuli involved the same pictures, but the pictures were paired with sentences featuring expressions depicting a value range whose boundary does not coincide with n:

1. *at most* $n + 1$
2. *fewer than* $n + 2$
3. *at least* $n - 1$
4. *more than* $n - 2$
5. *at most* $n - 2$
6. *fewer than* $n - 1$
7. *at least* $n + 2$
8. *more than* $n + 1$

Again, the first four were 'mathematically true' and the second four were 'mathematically false', and again, the second four were systematically judged as false by the informants. But in this case, there was one 'mathematically true' condition, namely *at most* $n+1$, where the sentences were judged as false at a significant rate (76%). A third experiment was carried out pitting *at most* n against *at most* $n + 1$ against each other directly, and again the *at most* $n + 1$ condition got a low rate of 'true' responses (44%), while *at most* n remained at ceiling.

So there is a particular difficulty that arises when the subject is looking at a picture of n objects and asked to judge whether it is true or false that there are *at most* $N + 1$ objects in the picture. This is supported by qualitative comments given by the participants (not reported by C&B). Among the comments that participants gave for such cases were the following:

- Looking at a picture of five Buddhas, asked to judge whether there are at most six, one participant wrote: "At most, there are 5," and marked it "false".
- Looking at a picture of three candles, asked to judge whether there are at most four, one participant wrote, "Technically true, but a very weird thing to say," and marked it "true".
- Looking at five mugs, asked to judge whether there are at most six, one participant wrote, "This one is hard. I'm marking it true, but it's super-weird."

Again, with respect to a picture containing n objects, it is unproblematic to judge a sentence with *at most* n as true (Experiment 1), but English speakers resist judging a corresponding sentence with *at most* $n + 1$ as true (Experiments 2 and 3).[2]

This pattern of results cannot be explained purely on the basis of ignorance implicatures, because there are other cases where an ignorance implicature arises and participants have no problem judging the sentence to be true. For example, *There are at least 4 butterflies* signals epistemic uncertainty with respect to the possibility that there are five butterflies, but participants nearly unanimously judge the sentence to be true even when it is clear exactly how many butterflies there are. Something beyond ignorance implicatures is needed for this case.

C&B explain this pattern with the help of the concept of *highlighting* in inquisitive semantics and a new Gricean maxim, the Maxim of Depictive Sincerity. The idea is that *at most* n highlights the possibility that there are n objects of the relevant kind, as depicted in

[2]Spychalska (2013) found this effect as well using sentences like 'At most three stars are red' and corresponding pictures.

Figure 4, for an example where exactly four objects of the relevant kind are pictured (e.g. four butterflies).

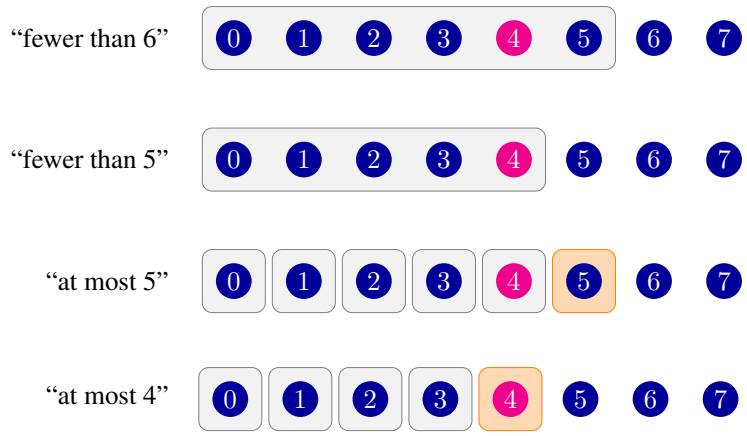


Figure 4: C&B's highlighting analysis

The red color on '4' marks that possibility as the actual one. The orange color indicates the highlighted possibility. For *at most five*, the highlighted possibility is distinct from the actual possibility, but for *at most four*, the highlighted possibility is the actual possibility. The Maxim of Depictive Sincerity requires that the speaker find the highlighted possibility epistemically accessible, so it is violated in the case of *at most five* but not in the case of *at most four*. The idea is that depictive sincerity implicatures are particularly strong, so strong that they can case participants to judge true sentences as false when they are violated.

Regarding methodology, C&B make the following conjecture. The picture-verification task methodology draws a line between two classes of implicatures: ignorance implicatures, which disappear in the picture-verification task, and depictive sincerity implicatures, which can be detected. In other words, picture scenario true/false judgement tasks can cut through certain types of pragmatic infelicity that complicate the interpretation of inference judgments, but even such true/false tasks are not impervious to particularly strong pragmatic requirements.

3 Towards a more direct comparison

Before C&B's methodological conjecture can be elevated to the status of a scientific finding, quite a bit more work needs to be done, as there were some potentially important differences between Geurts et al.'s paradigm and C&B's, besides truth judgments vs. validity judgments. C&B's participants were English speakers while Geurts et al.'s participants were Dutch speakers. The nature of the sentences was different; C&B used presentational *there* constructions (e.g. 'There are three bananas in the picture'), while Geurts et al. had the superlative modifier in object position (as in 'Berta drank three beers'). C&B's experiment was online, included pictures, and involved participants recruited via Mechanical

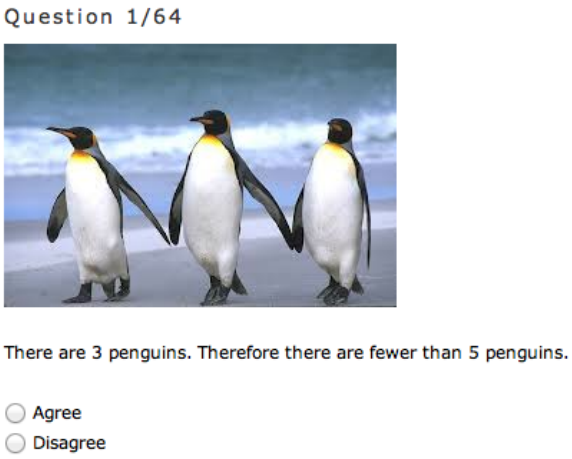


Figure 5: Example stimulus for validity judgment task (with pictures)

Turk. Geurts et al.'s participants were students who completed a written questionnaire not including pictures with paper and pencil. The question we investigated here was what happens if we eliminate these differences.

In service of a more controlled investigation of the effect of task, a validity judgment task using the stimuli from C&B's picture verification experiments was carried out. An example stimulus is shown in Figure 5. In one version of the experiment, the argument to be evaluated was accompanied by a picture, as in the picture verification task (Validity + Pictures). In the other version (Validity), there was no accompanying picture, as in Geurts et al.'s (2010) task. The stimuli in this condition are just as in Figure 5 except that no picture is shown. Varying whether there is an accompanying picture allows us to investigate whether the mere presence of a picture is responsible for the previously observed contrast between validity judgment tasks and picture verification taks. If so, then the version of the validity judgment task with a picture should yield similar results to the picture-verification task and the one without the picture should replicate the Geurts et al. (2010) pattern.

In total, four versions of the validity experiment were run. In addition to whether or not there was an accompanying picture, it was varied whether the modified numeral depicted a range ending on the numeral (as in C&B's Experiment 1), or off the numeral by one (as in C&B's Experiment 2). We will refer to these conditions as 'On Boundary' and 'Off Boundary', respectively. In all versions of the validity experiment, the arguments to be judged were of the form *There are* n *[nouns]. Therefore, there are* ___ *[nouns]*. In the true sentences of the 'On Boundary' condition, the blank was filled in *more than* $n-1$, *less than* $n+1$, *at least* n, or *at most* n. In the 'Off Boundary' condition, the blank was filled in by *more than* $n-2$, *less than* $n+2$, *at least* $n-1$, and *at most* $n+1$.

The results for all four experiments, alongside the results from C&B's picture verification experiments 1 and 2, are shown in Figure 6.[3] Visual inspection of the graphs shows

[3]Only the results for the 'mathematically true' stimuli are shown; the judgments for the 'mathematically false' stimuli were extremely close to 100% 'false' in all conditions.

that the pattern of responses turned out to be roughly the same across tasks. In particular, in the 'On Boundary' conditions depicted on the left, all responses are at ceiling, regardless of which modified numeral is used, in both variants of the validity judgment task as well as in C&B's picture verification task. However, the validity experiments did pick up on depictive sincerity implicatures, just like the picture verification experiment. This is shown by the graphs on the right, where the mean for *at most* $n + 1$ is consistently substantially lower than the others.

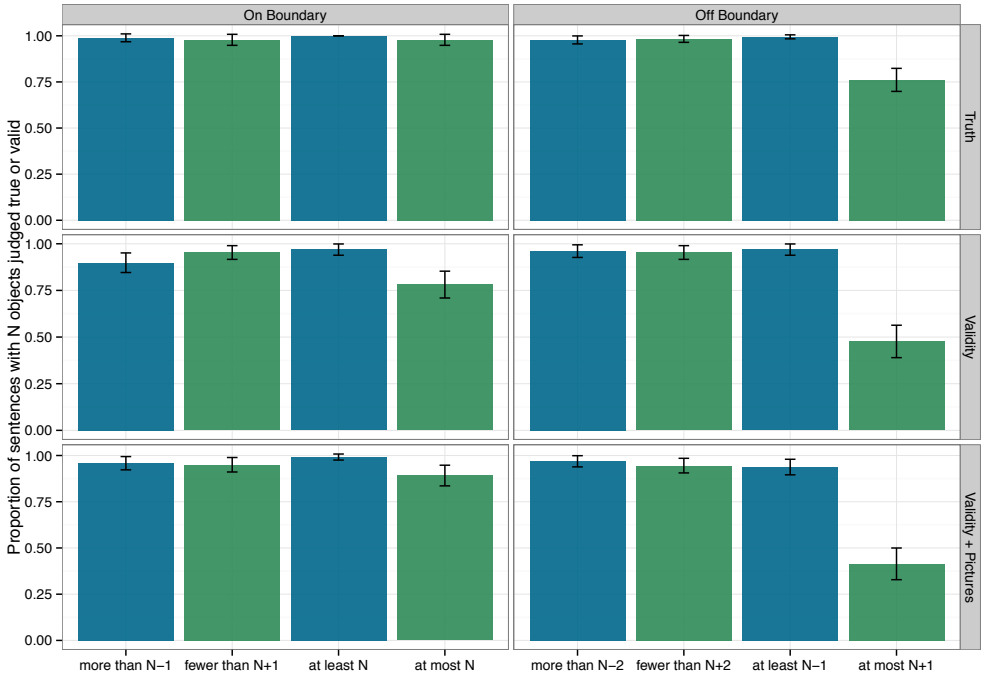


Figure 6: Results from picture verification ('truth') and validity experiments (picture verification results reported in Coppock & Brochhagen 2013a). The vertical axis of each graph shows the proportion of cases judged true or valid. Error bars show a 95% confidence interval.

These visual impressions are supported by the statistical analysis shown in Table 1, a mixed-effects regression model with random effects for subject and item, and fixed effects for the following four factors and their interactions.

- Modifier type (comparative vs. superlative)
- Upper- vs. lower-bounding (where *more than* and *at least* are lower-bounding and *fewer than* and *at most* are upper-bounding)
- 'On boundary' (e.g. *at most* n, as in C&B's Experiment 1) vs. 'off boundary (e.g. *at most* $n + 1$, as in C&B's Experiment 2)
- Task (picture verification vs. validity)

Presence vs. absence of pictures is not included in the model because it correlates very strongly with task (as the picture verification task always has a picture), and this factor is not significant either on its own or in combination with other factors for the dataset restricted

	Coeff.	Std. Err	Sig.
(Intercept)	0.98	(0.02)	***
modtype=sup	0.02	(0.03)	
bounding=upper	0.01	(0.03)	
cut=on	0.01	(0.04)	
task=validity	−0.01	(0.03)	
modtype=sup × bounding=upper	−0.24	(0.04)	***
modtype=sup × cut=on	−0.01	(0.04)	
bounding=upper × cut=on	−0.02	(0.04)	
modtype=sup × task=validity	−0.03	(0.03)	
bounding=upper × task=validity	−0.02	(0.03)	
cut=on × task=validity	−0.05	(0.05)	
modtype=sup × bounding=upper × cut=on	0.23	(0.06)	***
modtype=sup × bounding=upper × task=validity	−0.25	(0.05)	***
modtype=sup × cut=on × task=validity	0.07	(0.05)	
bounding=upper × cut=on × task=validity	0.06	(0.05)	
modtype=sup × bounding=upper × cut=on × task=validity	0.09	(0.08)	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1: Coefficients, standard errors, and significance levels for fixed effects in `lmer` model

to the validity judgments.

The model in Table 1 is constructed using the `lmer` function of `lme4` package in `R` with the formula:

```
result ~ modtype*bounding*cut*task + (1|subj) + (1|noun)
```

where `modtype` is modifier type, `bounding` refers to upper- vs. lower-bounding, `cut` refers to where the numerical description makes a cut between true and false, and `task` is either 'picture verification' or 'validity'. Significance levels are based on t-tests as estimated by the `lmerTest` package.

The table shows significant effects for three interaction factors.

1. First, *at most* is special, as shown by the significance of the factor 'modtype=sup × bounding=upper', which had a relatively large negative coefficent (so *at most* is generally difficult, even controlling for interactions with other factors).
2. Furthermore, *at most* $n + 1$ is different from *at most* n, as shown by the effect of the interaction factor 'modtype=sup × bounding=upper × cut=on'. The direction of this factor indicates that, as can be seen clearly in the graphs, *at most* $n + 1$ is harder than *at most* n. Note that the model includes an interaction with task, so this factor is not driven purely by the picture verification task; it also plays a role for validity judgments.
3. The effect of *at most* turned out to be more pronounced in the validity task, hence significance for the interaction factor 'modtype=sup × bounding=upper × task=validity'.

Crucially, there was no main effect of modifier type (comparative vs. superlative), and nor

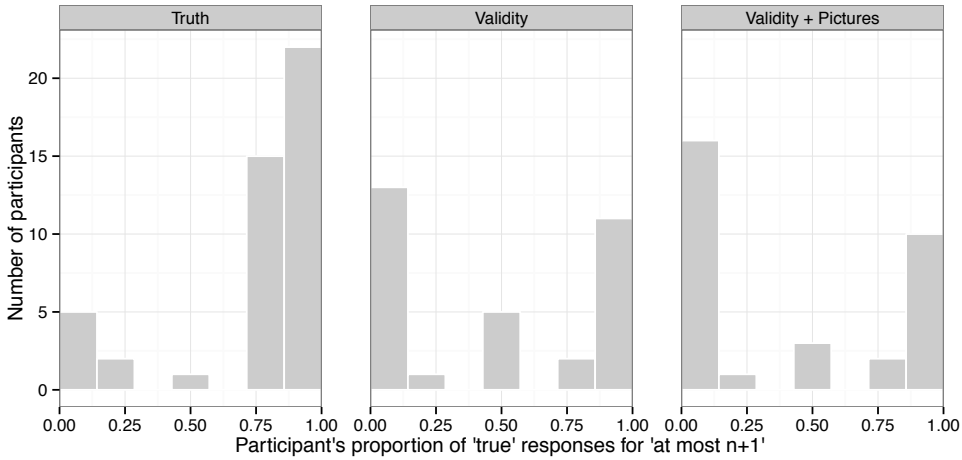


Figure 7: Histogram of responses for 'at most $n + 1$'

did this factor have substantially more of an effect in the validity task; there was *no* interaction between task and modifier type. This constitutes a failure to replicate Geurts et al. (2010), where acceptance rates for both *at least* n and *at most* n were substantially lower than for their comparative counterparts *more than* $n - 1$ and *fewer than* $n + 2$. This also goes against C&B's hypothesis that validity judgment tasks are more sensitive to ignorance implicatures than picture verification tasks.

Note that the validity judgment task using the Coppock & Brochhagen items was not completely immune to pragmatic oddity, however. It did pick up depictive sincerity violations. As in the picture verification task, the inference from *n* to *at most* $n + 1$ is accepted about half of the time. Inspection of the histograms in Figure 7 shows that this is the result of a bimodal distribution among the participants, where some consistently judge the relevant inferences as invalid, and some judge them as valid. So there appear to be two kinds of speakers: those that treat depictive sincerity violations as grounds for judging an inference to be invalid, and those that adhere to more 'mathematical' intuitions.[4]

4 Conclusion

The methodological conclusions we may tentatively draw from these investigations as follows. Validity judgment tasks may be more sensitive to ignorance implicatures than picture verification tasks under some conditions, but validity judgments do not robustly pick up on ignorance implicatures, so they cannot be relied upon for that, whether or not there is an accompanying picture. Validity judgments, like truth value judgments, *are* sensitive to depictive sincerity implicatures. However, there appear two classes of individuals: Those who treat depictive sincerity violations as grounds for invalidity, and those that adhere to more 'mathematical' intuitions.

[4]The distribution of responses in the picture verification task was not bimodal; this difference may be interesting to study further.

Much more work needs to be done to determine the conditions under which Geurts et al.'s (2010) findings replicate. Note that Cummins & Katsos (2010) *did* replicate Geurts et al.'s (2010) findings in a validity judgment task (called 'implication judgment task' there), using 15 native English speakers in Cambridge. The inference from *three* to *at least three* was accepted 62% of the time, whereas the inference from *three* to *more than two* was accepted 100% of the time. The inference from *three* to *at most three* was accepted 42% of the time, and the inference from *three* to *fewer than four* was accepted 84% of the time. These numbers roughly correspond to Geurts et al.'s (2010) findings. So the difference of language seems an unlikely culprit. It may have to do with the way the question is phrased, however. The experimental paradigm used there is described as follows (p. 289): "They were informed that they would see a series of pages, each with two sentences written on them, and that they should circle the answer 'yes' if the first sentence implied the second and 'no' if it did not." There are many additional possible explanations for the present failure to replicate, as there are many differences between this validity experiment and previous ones. The lack of effect may be due to the syntactic position in which the superlative modifier was places (pivot of an existential construction vs. object of a transitive verb), the way that the question was phrased, the mix of experimental items and fillers, or a difference in the sample of participants. These factors should be investigated systematically in future research.

References

Biezma, María. 2013. Only one *at least*: Refining the role of discourse in building alternatives. In *University of Pennsylvania Working Papers in Linguistics* 19, 11–19. Penn Linguistics Club.

Büring, Daniel. 2008. The least *at least* can do. In Charles B. Chang & Hannah J. Haynie (eds.), *26th West Coast Conference on Formal Linguistics*, 114–120. Somerville, MA: Cascadilla Press.

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2008. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In Claudia Maienborn, Klaus von Heusinger & Paul Portner (eds.), *Semantics: An international handbook of natural language meaning*, Berlin: Mouton de Gruyter.

Cohen, Ariel & Manfred Krifka. 2011. Superlative quantifiers as modifiers of meta-speech acts. In Barbara H. Partee, Michael Glanzberg & Jurgis Skilters (eds.), *The Baltic International Yearbook of Cognition, Logic and Communication*, vol. 6, 1–56. New Prairie Press.

Coppock, Elizabeth & Thomas Brochhagen. 2013a. Diagnosing truth, interactive sincerity, and depictive sincerity. In *Proceedings of SALT 23*, eLanguage.

Coppock, Elizabeth & Thomas Brochhagen. 2013b. Raising and resolving issues with scalar modifiers. *Semantics & Pragmatics* 6(3). 1–57.

Cummins, Chris & Napoleon Katsos. 2010. Comparative and superlative quantifiers: Pragmatic effects of comparison type. *Journal of Semantics* 27. 271–305.

Geurts, Bart, Napoleon Katsos, Chris Cummins, Jonas Moons & Leo Noordman. 2010. Scalar quantifiers: Logic, acquisition, and processing. *Language and Cognitive Processes* 25.

Geurts, Bart & Rick Nouwen. 2007. *At least* et al.: The semantics of scalar modifiers. *Language* 83. 533–559.

Groenendijk, Jeroen & Floris Roelofsen. 2009. Inquisitive semantics and pragmatics. Presented at the Workshop on Language, Communication, and Rational Agency at Stanford, May 2009, available via `www.illc.uva.nl/inquisitivesemantics`.

Kaplan, David. 1999. The meaning of *ouch* and *oops*. Lecture presented at the University of California at Berkeley.

Kennedy, Christopher. 2015. A "de-Fregean" semantics (and neo-Gricean pragmatics) for modified and unmodified numerals. *Semantics and Pragmatics* 8. 1–44.

Mayr, Clemens. 2013. Implicatures of modified numerals. In Ivano Caponigro & Carlo Cecchetto (eds.), *From grammar to meaning: The spontaneous logicality of language*, 139–171. Cambridge: Cambridge University Press.

McNabb, Yaron & Doris Penka. 2014. The processing cost of interpreting superlative modifiers and modals. In Judith Degen, Michael Franke & Noah Goodman (eds.), *Proceedings of Formal & Experimental Pragmatics 2014*, 29–35.

Schwarz, Bernhard. 2013. At least and quantity implicature: Choices and consequences. In Maria Aloni, Michael Franke & Floris Roelofsen (eds.), *Proceedings of the 19th amsterdam colloquium*, 187–194.

Schwarz, Bernhard. to appear. Consistency preservation in quantity implicature: the case of *at least*. *Semantics & Pragmatics* .

Schwarz, Bernhard & Junko Shimoyama. 2011. Negative islands and obviation by *wa* in Japanese degree questions. In Nan Li & David Lutz (eds.), *Proceedings of SALT 20*, 702–719. eLanguage.

Spychalska, Maria. 2013. Pragmatic effects in processing superlative and comparative quantifiers: epistemic-algorithmic approach. Slide presentation (available on author's website).