What is at-issueness? An experimental comparison of diagnostics

Word count (including references, excluding abstract and supplements): 11,924

Abstract At-issueness is a central concept in theoretical semantics and pragmatics, but there is no consensus about how it should be defined or diagnosed (e.g., Tonhauser 2012; Snider 2017b; 2018; Tonhauser et al. 2018; Koev 2018; Faller 2019; Korotkova 2020). We present the results of six experiments designed to investigate whether five diagnostics for at-issueness yield consistent results. Our findings reveal substantial differences between diagnostics, indicating that they are not interchangeable and that the choice and implementation of diagnostics matter for empirical generalizations about at-issueness. Diagnostics that measure the at-issueness of contents embedded in questions show the greatest differentiation to me across expressions, indicating that speech act embedding plays a key role.**JT: key role in what? in general, how we assessed the diagnostics (differentiation) and how they differ (declarative vs. interrogative, response tasks) needs to be set up better, so that this sentence doesn't come so out of the blue** However, the theoretical divide between QUD-based and assertion-based diagnostics assumed in previous literature **JT: this also comes out of the blue** does not appear to be the primary source of divergence. **JT: the word "divergence" is also hard to interpret here because we haven't said anything yet about how we'll evaluate the diagnostics and what the results were** Instead, we argue that the crucial difference lies in how questions and assertions relate at-issue and not-at-issue content to speaker commitments. **JT: this sentence belongs with the first sentence above, about declarative vs. interrogative**

Keywords: at-issueness, diagnostics, experimental pragmatics

1 Introduction

At-issueness is a key concept in theoretical semantics and pragmatics,[1] used in the analysis of various phenomena, including presupposition, conventional implicature, evidentials, expressives, and co-speech gestures (e.g., Karttunen & Peters 1979; Horton & Hirst 1988; Abbott 2000; Faller 2002; 2019; Potts 2005; Simons et al. 2010; Lee 2011; Tonhauser et al. 2018; Esipova 2019; Korotkova 2020; Barnes & Ebert 2023; Chen 2024; Scontras & Tonhauser 2025). It is generally understood as distinguishing the main point of an utterance (at-issue content) from propositions conveying background information (not-at-issue content), but there is no consensus how at-issueness should be diagnosed, and competing definitions reflect different assumptions about its underlying nature (e.g., Tonhauser 2012; Snider 2017a; b; 2018; Tonhauser et al. 2018; Koev 2018; Faller 2019; Esipova 2019; Korotkova 2020). Consequently, empirical claims about whether a given expression contributes at-issue or not-at-issue content may be relative to specific diagnostics, and different diagnostics may not target the same underlying phenomenon.

Four commonly used at-issueness diagnostics are illustrated in (1–4) for sentence-medial appositive non-restrictive relative clauses (NRRCs), which are typically taken to contribute not-at-issue content (Potts 2005). Accordingly, participants are expected to: give low question-answer-match ratings under the QUD diagnostic (1) (Amaral et al. 2007; Lee 2011; Tonhauser 2012; Chen

[1] Potts (2015) credits the term to Bill Ladusaw, "who began using it informally in his UCSC undergraduate classes in 1985" (p. 2).

2024); judge that the speaker is not asking about the NRRC content under the 'asking-whether' diagnostic (2) (Tonhauser et al. 2018; Solstad & Bott 2024; Degen & Tonhauser 2025); give low naturalness ratings under the direct-dissent diagnostic (3), (Faller 2002; 2006; Papafragou 2006; Amaral et al. 2007; Murray 2010; 2014; AnderBois et al. 2010; 2015; Tonhauser 2012; Syrett & Koev 2015); and prefer signalling agreement with the previous assertion (i.e. a *yes*-response) when rejecting the NRRC content under the 'yes, but' diagnostic (4), (Xue & Onea 2011; Cummins et al. 2013; Destruel et al. 2015).

(1) QUD diagnostic

Nora: *What did Greg buy?*
Leo: *Greg, who bought a new car, is envied by his neighbor.*

Question to participants: How well does Leo's response fit Nora's question?

(2) 'asking whether' diagnostic

Nora: *Is Greg, who bought a new car, envied by his neighbor?*

Question to participants: Is Nora asking whether Greg bought a new car?

(3) Direct-dissent diagnostic

Nora: *Greg, who bought a new car, is envied by his neighbor.*
Leo: *No, that's not true, he didn't buy a new car.*

Question to participants: How natural is Leo's rejection of Nora's utterance?

(4) 'yes, but' diagnostic

Nora: *Greg, who bought a new car, is envied by his neighbor.*
Leo: *Yes, but he didn't buy a new car.* /
Yes, and he didn't buy a new car. /
No, he didn't buy a new car.

Task for participants: Choose the response that sounds best.

The assumptions underlying these diagnostics can be organized around three properties of at-issue content identified in Tonhauser (2012) as motivating diagnostic strategies:

(i) at-issue content addresses the question under discussion (see also Amaral et al. 2007; Simons et al. 2010),
(ii) the at-issue content of questions determines the relevant set of alternatives (see also Tonhauser et al. 2018), and
(iii) at-issue content can be directly assented or dissented with (see also Shanon 1976; Faller 2002; 2006; Murray 2010).

The QUD diagnostic (1) (Lee 2011; Tonhauser 2012; Chen 2024) targets property (i), based on the view that discourse is structured around addressing a *Question Under Discussion* (QUD) (Roberts 1996; Ginzburg 1996) and that at-issue content is the part of an utterance intended to address that question (Amaral et al. 2007; Simons et al. 2010). The diagnostic, therefore, assumes that only at-issue content can felicitously address an established QUD. The diagnostic presents the target content in a response to a question about that content, and question-answer match ratings are expected to be high when the target content can be construed as at-issue and low otherwise.

The 'asking whether' diagnostic (2) (e.g., Tonhauser et al. 2018; Solstad & Bott 2024; Degen & Tonhauser 2025) targets property (ii), testing whether the target content is interpreted as determining the question alternatives. Participants judge whether a speaker, who asks a polar question containing the target content, is asking whether that content is true. Such 'asking whether' judgments are expected to be high when the target content is at-issue and low otherwise.

The direct-dissent diagnostic (3) (e.g., Faller 2002; 2006; Papafragou 2006; Amaral et al. 2007; Murray 2010; 2014; AnderBois et al. 2010; 2015; Tonhauser 2012; Syrett & Koev 2015) and the 'yes, but' diagnostic (4) (Xue & Onea 2011; Cummins et al. 2013; Destruel et al. 2015) target property (iii), resting on the idea that only at-issue content can be directly affirmed or denied, whereas rejecting not-at-issue content requires more indirect discourse moves (Faller 2002; 2006; Potts 2005; Papafragou 2006; Amaral et al. 2007). The direct-dissent diagnostic assumes that directly dissenting with the target content is judged natural if the content is at-issue and unnatural otherwise. Relatedly, the 'yes, but' diagnostic tests whether a target content can be denied while responding *yes* to the assertion as a whole. Originally developed to diagnose pragmatic inferences that are not semantic entailments (Onea & Beaver 2009), it has been adopted as an at-issueness diagnostic under the assumption that *yes, but* responses provide an indirect denial suitable for dissenting with not-at-issue content (Xue & Onea 2011).

Although these diagnostics are widely used, there is no consensus on whether the properties they target correspond to a single underlying notion of at-issueness. Tonhauser (2012) argued that diagnostics targeting properties (i)–(iii) can be understood as probing at-issueness relative to the QUD, adopting the definition in (5), where *m* may be a proposition or a question denotation.

(5) QUD at-issueness: (based on Simons et al. 2010, p. 323)
A content *m* is at-issue in a context *c* iff
a. *m* is relevant[2] to the QUD in *c*, and
b. the speaker has a recognizable intention to address the QUD via *m*.

While this definition aligns naturally with property (i), it is less clear whether all diagnostics, in particular those targeting property (iii), track this QUD-based notion.

Evidence for differences between diagnostics has been reported, for instance for appositive NRRCs: While appositive NRRCs are widely taken to contribute not-at-issue content, a forced-choice continuation study by Syrett & Koev (2015) found that sentence-final appositive NRRCs pattern as more at-issue than medial ones, under a version of the direct-dissent diagnostic. Snider (2017b) argued that the direct-dissent diagnostic is sensitive to this medial/final distinction, but diagnostics targeting properties (i) or (ii) are not. Similar dissociations have been reported for sentence-final slifting parentheticals (Koev 2018) and evidential inferences (Korotkova 2020).

These kinds of observations have motivated the view that at-issueness is not a uniform notion (Koev 2018). In particular, property (iii) has been linked to an assertion-based notion of at-issueness, grounded in the treatment of assertions as proposals to update the common ground (Stalnaker 1978; Clark & Schaefer 1989; Ginzburg 1995; Farkas & Bruce 2010). On this view, the at-issue content of an assertion is the proposition proposed to be added to the common ground, whereas not-at-issue content is either presupposed (already entailed in the common ground; Stalnaker 1973; 2002), or newly imposed without being proposed (Murray 2010; 2014; AnderBois et al. 2010; 2015). This notion is defined as follows:[3]

(6) Proposal at-issueness: (Koev 2013, pp. 50–51)
A proposition *p* is at-issue in a context *c* iff
a. *p* is a proposal in *c* and
b. *p* has not been accepted or rejected in *c*.

[2] Relevance to the QUD can be defined in terms of contextual entailment of a partial or complete answer (Roberts 2012; Simons et al. 2010), or based on inferences about the probability of QUD-answers (Büring 2003; Beaver & Clark 2008).

[3] Further definitions of at-issueness proposed in the literature include analyses in terms of restrictive vs. non-restrictive modification (Esipova 2019; 2021), and coherence-based discourse structure (Hunter & Asher 2016; Jasinskaja 2016). Clarifying the theoretical relations among the various definitions of at-issueness proposed in the literature remains an important question for future research.

An alternative line of work argues that diagnostics targeting property (iii) do not diagnose at-issueness at all, but rather the availability of the target content for propositional anaphora, such as response particles (*yes/no*) or propositional proforms (e.g., *that* in *that's not true*) (Snider 2017b; a; 2018; Korotkova 2020). This perspective is compatible with analyses that suggest that QUD-based and proposal-based notions ultimately converge: AnderBois et al. (2015), building on Farkas & Bruce (2010), argue that proposing content for the common ground is intrinsically linked to managing the QUD, as reflected in parallels between assertions and polar questions. Similarly, Faller (2019) analyzes assent and dissent as discourse moves that address the QUD.

The literature thus raises the question whether the tension between diagnostics targeting properties (i) and (iii) reflect genuinely distinct notions of at-issueness, or arises from differences in how diagnostics interact with other discourse factors, like the anaphoric availability of the target content. Property (ii) can be seen as a forward-looking counterpart of property (i) (Snider 2017b): while property (i) concerns addressing an established QUD, property (ii) concerns determining which alternatives a question introduces. Under standard assumptions that explicit questions introduce or update the QUD (Roberts 1996; Ginzburg 1996), property (ii) aligns naturally with the QUD-based definition in (5), but is incompatible with assertion-based definitions, which apply only to declaratives.

In sum, existing diagnostics differ both in the properties they target and in the theoretical notions they are taken to diagnose. Table 1 summarizes how the four diagnostics discussed here have been mapped to different notions of at-issueness: While Tonhauser (2012), AnderBois et al. (2015) and Faller (2019) ultimately assume that at-issueness diagnostics target a uniform, QUD-based notion, Snider (2017a; b; 2018) and Korotkova (2020) argue that diagnostics targeting property (iii) do not diagnose at-issueness of content, but its availability for propositional anaphora; and Koev (2018) argues for two distinct notions, corresponding to QUD-based and proposal-based at-issueness.

<table>
<tr><th rowspan="2">Diagnostic</th><th rowspan="2">Property targeted by diagnostic</th><th colspan="3">Underlying notion assumed to correspond to property</th></tr>
<tr><th>Tonhauser 2012, AnderBois et al. 2015, Faller 2019</th><th>Snider 2017a, Korotkova 2020</th><th>Koev 2018</th></tr>
<tr><td>(1) QUD diagnostic</td><td>(i) addressing QUD</td><td rowspan="4">QUD at-issueness</td><td rowspan="2">QUD at-issueness</td><td>QUD at-issueness</td></tr>
<tr><td>(2) Asking whether</td><td>(ii) determining alternatives</td><td>–</td></tr>
<tr><td>(3) Direct dissent
(4) 'Yes, but'</td><td>(iii) direct dissent/ assent</td><td>Anaphoric availability</td><td>Proposal at-issueness</td></tr>
</table>

Table 1: Mapping between at-issueness diagnostics, the properties they target (Tonhauser 2012), and the underlying theoretical notions assumed in different theoretical approaches.

Although the empirical differences between at-issueness diagnostics discussed above have been documented across a range of constructions, but to date they have not been evaluated in a systematic experimental comparison applying multiple diagnostics to the same target contents. Together with ongoing theoretical disagreement about what at-issueness is, this raises two central questions:

i. Do the various diagnostics yield consistent results?
ii. Do the existing definitions of at-issueness target the same underlying phenomenon?

Initial research on these questions has yielded mixed answers.

First, as noted above, Syrett & Koev (2015) reported experimental evidence for a positional effect for appositive NRRCs under a version of the direct-dissent diagnostic. Since this medial/final

distinction has been argued to assent/dissent-based diagnostics, a key question is whether the same contrast is reproduced by other at-issueness diagnostics when applied to identical materials.

Beyond appositives, Tonhauser et al. 2018 compared at-issueness measures for a range of English expressions including sentence-medial appositives and the clause-embedding predicates *discover, know,* and *annoyed*, using two diagnostics: the question-based 'asking whether' diagnostic and the 'are you sure?' (p. 526ff).[4] They found that contents received lower overall ratings under the 'asking whether' diagnostic, while the 'are you sure?' diagnostic showed greater differentiation across expressions. Despite these differences, ratings from both diagnostics were correlated, suggesting that they may nonetheless be sensitive to a shared underlying phenomenon.

These findings suggest that diagnostics may differ in how strongly they differentiate among contents. This raises another key question for comparing diagnostics: Do diagnostics differ on whether they reproduce fine-grained by-content differences reported in the literature? We address this by examining the complements of clause-embedding predicates, for which Degen & Tonhauser (2025) report fine-grained differences across 20 English predicates under the 'asking whether' diagnostic. Figure 1 shows the mean 'asking whether' ratings by predicate in their study. The question is whether comparable patterns of differentiation emerge across diagnostics when applied to the content of the complements of (some of) these predicates.

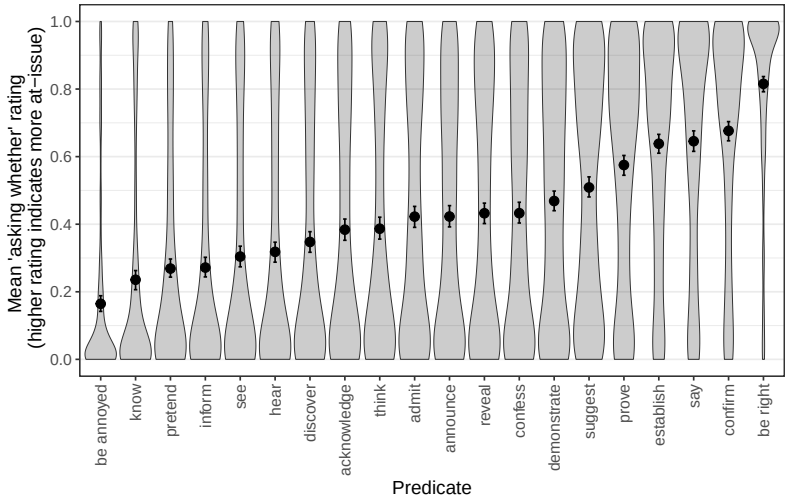


Figure 1: Mean 'asking whether' ratings for the contents of the clausal complements of 20 clause-embedding predicates, from Degen & Tonhauser 2025. Error bars indicate 95% bootstrapped confidence intervals. Violin plots indicate the distribution of individual ratings.

This paper takes a first step toward addressing the above questions (i+ii) through by systematically assessing five diagnostics of at-issueness across six experiments.

Our study asks whether the diagnostics give rise to a consistent pattern of results when applied to the same target contents by comparing the diagnostics on whether they reproduce contrasts reported in prior work, specifically, by examining whether different diagnostics converge on the medial versus sentence-final distinction for appositive NRRCs, and fine-grained by-predicate differences among the complements of clause-embedding predicates.

[4] The 'are you sure?' diagnosic presents participants with dialogues like (i) (Tonhauser et al. 2018, p. 519) and collects ratings about whether Fred ansered Carla's question (see also the 'really' test of Shanon 1976). This diagnostic cannot clearly be aligned with the three properties of at-issue content.

(i) Fred: Marthas new car, a BMW, was expensive.
Carla: Are you sure?
Fred: Yes, I am sure that Marthas new car is a BMW.

Experiments 1–4 apply the four diagnostics to the same set of seven contents in English, including appositive NRRCs and the complements of selected clause-embedding predicates. While none of these four experiments observed the positional effect for NRRCs found by Syrett & Koev (2015), only Exp. 2 ('asking whether') comes close to observing the differences between the clause-embedding predicates observed in Degen & Tonhauser (2025), suggesting that diagnostics do vary in whether they detect at-issueness differences between particular expressions.

The 'asking whether' diagnostic, as implemented in Exp. 2, which showed the greatest differentiation between contents, is the only diagnostic which asked participants to judge the intentions of the speaker rather judgments related to acceptability, and it is the only diagnostics where the target content is embedded in a polar question.

To test whether the greater differentiation under the 'asking whether' diagnostic was due to the interrogative embedding or from the response task itself, we compared the 'asking whether' diagnostic (Exp. 5) with a diagnostic that also embeds content in a polar question, but elicits acceptability judgments for a direct response (Exp. 6), like in the direct-dissent diagnostic, using the same 20 clause-embedding predicates from Degen & Tonhauser (2025). The two experiments yielded highly correlated results, indicating that the observed increased differentiation is driven by the interrogative embedding, not the response task.

Our results, therefore, provide evidence that the speech act used to present the target content plays an important role: embedding the target content in questions yield results with greater by-content differentiation than embedding it in assertions.

2 Experiments 1-4

To compare the results of at-issueness diagnostics, we conducted four experiments that each measured at-issueness with a different diagnostic, namely the QUD diagnostic (Exp. 1), the 'asking whether' diagnostic (Exp. 2), the direct-dissent diagnostic (Exp. 3) and the 'yes, but' diagnostic (Exp. 4).[5] Each experiment tested the same manipulation,**JT: do experiments test manipulations?** comparing the propositional contents**JT: see above on 'propositional'?** of the seven types of expressions (contents)**JT: i don't understand this phrasing: is "know" a type of expression? is the complement of 'know' a type of expression? and why does "content" occur in parentheses after "types of expressions"? we are comparing the at-issueness of seven contents, no?** illustrated in (7): the contents of sentence-medial and sentence-final NRRCs (7a)-(7b), as well as the contents of the clausal complements of *know, discover, confess, confirm* and *be right* (7c)-(7g). Contents were randomly paired with items from a set of items shared across all four experiments (see example pairings in (7)).**JT: this last sentence belongs in the methods section**

(7) a. Content of sentence-medial NRRC
Lucy, who broke the plate, apologized. ⇝ Lucy broke the plate

b. Content of sentence-final NRRC
The police found Jack, who saw the murder. ⇝ Jack saw the murder

c. Content of the clausal complement of *know*
Ann knows that Raul cheated on his wife. ⇝ Raul cheated on his wife

d. Content of the clausal complement of *discover*
Mary discovered that Denny ate the last cupcake. ⇝ Denny ate the last cupcake

e. Content of the clausal complement of *be right*
Tom is right that Ann stole the money. ⇝ Ann stole the money

[5] See the Data accessibility statement for a link to the Github repository that provides access to the two experiments, the data, and the analysis scripts.

f. Content of the clausal complement of *confirm*
Harry confirmed that Greg bought a new car. ⇝ Greg bought a new car

g. Content of the clausal complement of *confess*
Lucy confessed that Dustin lost his key. ⇝ Dustin lost his keys

These seven contents were chosen because prior literature observed differences in at-issueness between two or more of these contents using a particular diagnostic for at-issueness. Specifically, as discussed in §1, Syrett & Koev 2015 observed differences between sentence-medial and -final NRRCs using a variant of the direct-dissent diagnostic, and Degen & Tonhauser 2025 observed differences between *know, discover, confess, confirm* and *be right* using the 'asking whether' diagnostic. Thus, comparing these seven contents across the four diagnostics in Exps. 1-4 will allow us to assess whether the differences that emerge from one diagnostic also emerge from others. In each experiment, participants read the stimuli and gave ratings corresponding to the diagnostics.

2.1 Methods

2.1.1 Participants

For each of the four experiments, we recruited 80 unique participants on Prolific. These participants had registered on the platform as living in the USA and as having English as their primary language. They had at least 50 previous submissions and an approval rate of at least 97%. Table 2 shows the age and gender distributions of the recruited participants.

	recruited	ages (mean age)	f/m/nb/dnd
Exp. 1 (QUD)	80	18-81 (43.8)	42/37/0/1
Exp. 2 (asking whether)	80	20-74 (38.5)	48/30/1/1
Exp. 3 (direct dissent)	80	18-77 (39.1)	50/28/1/1
Exp. 4 (yes, but)	80	19-67 (38.0)	48/30/2/0

Table 2: Information about the participants recruited in Exps. 1-4 (f = female, m = male, nb = nonbinary, dnd = did not disclose).

2.1.2 Materials and procedure

The four experiments measured the at-issueness of the seven contents in (7), each using a different diagnostic: the QUD diagnostic (Exp. 1), the 'asking whether' diagnostic (Exp. 2), the direct-dissent diagnostic (Exp. 3), and the 'yes, but' diagnostic (Exp. 4). (8) illustrates how each diagnostic was implemented using sentence-medial NRRCs (with the item 'Lucy broke the plate').**JT: this feels repetitive to examples (1)-(4) from the intro. perhaps just show Fig 2?**

In Exp. 1 (QUD diagnostic, (8a)), participants read a dialogue between two named speakers, where the first utters a constituent question (the presumed QUD) that is about the target content and the second responds with a declarative sentence that contributes the target content. In Exp. 2 ('asking whether' diagnostic, (8b)), participants read a polar question uttered by a named speaker, which itself contributes the target content. In Exp. 3 (direct-dissent diagnostic, (8c)), participants read a dialogue between two named speakers, where the first utters a declarative sentence with the target content and the second directly dissents with the target content. Finally, in Exp. 4 ('yes, but' diagnostic, (8d)), participants read a dialogue between two named speakers where the first utters a declarative sentence that contributes the target content and the second responds with one of two indirect dissent variants (*yes, but..*, *yes, and...*) or a direct dissent.

(8) Implementation of the diagnostics in Exps. 1-4

a. Exp. 1 (QUD diagnostic)
Nora: *What did Lucy break?*
Leo: *Lucy, who broke the plate, apologized.*

b. Exp. 2 ('asking whether' diagnostic)
Nora: *Did Lucy, who broke the plate, apologize?*

c. Exp. 3 (direct-dissent diagnostic)
Nora: *Lucy, who broke the plate, apologized.*
Leo: *No, she didn't break the plate.*

d. Exp. 4 ('yes, but' diagnostic)
Nora: *Lucy, who broke the plate, apologized.*
Nina: *Yes, but she didn't break the plate.*
Yes, and she didn't break the plate.
No, she didn't break the plate.

As shown in Fig. 2, the response options differed by diagnostic. In Exp. 1 (QUD diagnostic, panel (a)), participants rated how well the response fit the question on a slider marked 'totally doesn't fit' on one end (coded 0) and 'totally fits' on the other end (coded as 1). In Exp. 2 ('asking whether' diagnostic, panel (b)), participants judged whether the question was about the target content, using a slider marked 'no' on one end (coded as 0) and 'yes' on the other (coded as 1). In Exp. 3 (direct-dissent diagnostic, panel (c)), participants rated the naturalness or the direct dissent on a slider marked 'totally unnatural' (coded as 0) on one end and 'totally natural' on the other (coded as 1). Finally, in Exp. 4 ('yes, but' diagnostic, panel (d)), participants chose the response that sounded best; the two indirect dissents were coded as 0 and the direct one as 1. Across the four experiments, the responses were coded so that 1 meant that the content to be diagnosed was rated as at-issue and 0 as not-at-issue.

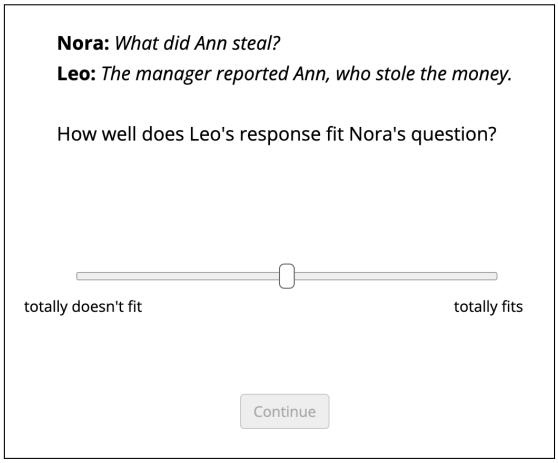
Each of the seven contents in (7) was combined with one of the seven items in (9) in each of the four experiments.

(9) a. Jack saw the murder.
b. Raul cheated on his wife.
c. Ann stole the money.
d. Danny ate the last cupcake.
e. Lucy broke the plate.
f. Dustin lost his key.
g. Greg bought a new car.

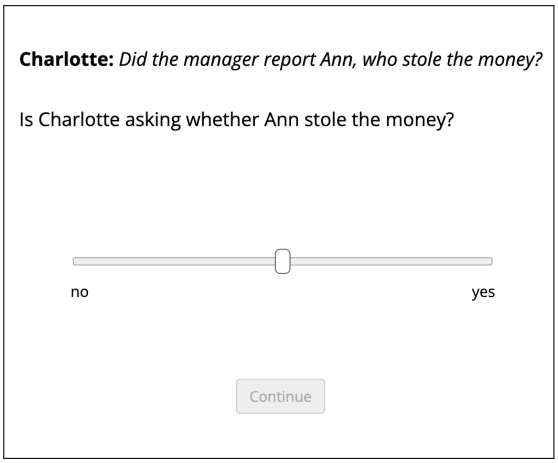
Each experiment included two control stimuli serving as attention checks: one was expected to receive a response at one end of the slider (Exps. 1-3) or a 'no' response (Exp. 4); the other control stimulus was expected to receive a response at the other end of the slider (Exps. 1-3) or a 'yes' response (Exp. 4). See Supplement A for the control stimuli used in Exps. 1-4.

In all four experiments, each participant's set of items was generated by randomly combining each of the seven contents in (7) with a unique content in (9). Participants completed a total of 9 trials, namely 7 target trials and the same 2 control trials. Trial order was randomized for each participant.

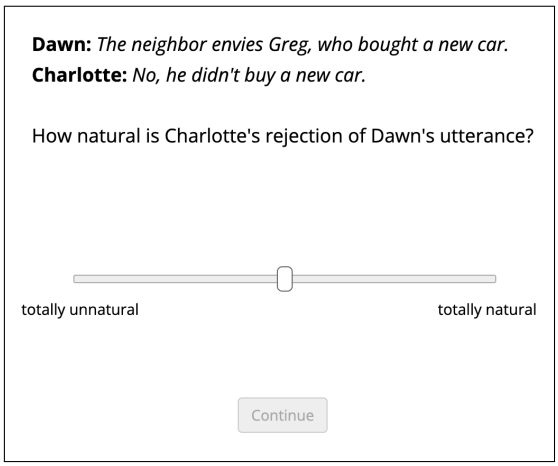
After completing the experiment, participants filled out a short optional demographic survey. To encourage truthful responses, participants were told that they would be paid no matter what answers they gave in the survey.



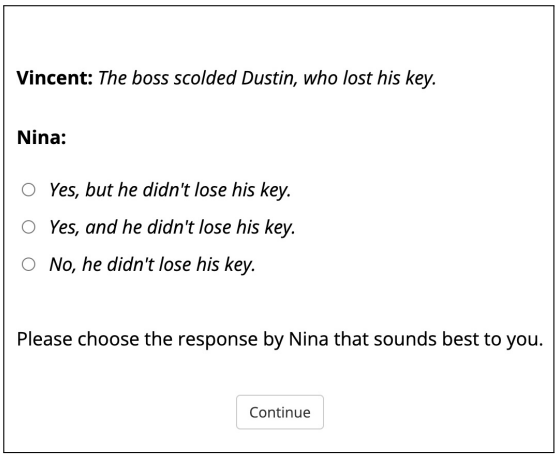
(a) Exp. 1: QUD diagnostic



(b) Exp. 2: 'asking whether' diagnostic



(c) Exp. 3: direct-dissent diagnostic



(d) Exp. 4: 'yes, but' diagnostic

Figure 2: Sample trials in (a) Exp. 1, (b) Exp. 2, (c) Exp. 3, and (d) Exp. 4.

2.1.3 Dependent measure and interpretation

Following prior work (e.g., Tonhauser et al. 2018), we interpret higher/lower mean ratings as indicating that content is more/less at-issue under a given diagnostic,**JT: this belongs in the methods section** with two caveats: First, we remain agnostic about whether at-issueness is an underlyingly binary or a gradient property (cf. Tonhauser et al. 2018; Barnes & Ebert 2023). If at-issueness is gradient, the extent to which a content is at-issue may be understood as the extent to which it is relevant to the QUD or the main assertion, in which case gradient mean ratings may be taken to reflect gradient relevance. If at-issueness is categorical, content is at-issue iff it addresses the QUD or assertive proposal, and not-at-issue otherwise; and gradient mean ratings could be attributed to uncertainty about what the QUD/proposal is. For example, our interpretation of a content in a given utterance being more/less at-issue may be interpreted as reflecting the frequency or ease with which a particular QUD is attributed to that utterance.**JT: is this really a caveat to the linking function?** Second, we use the term "at-issueness diagnostic" descriptively throughout

the paper, even though our findings may ultimately suggest that these diagnostics track distinct theoretical constructs. We return to this issue in the general discussion.

2.1.4 Data exclusion

We excluded the data of participants that were not self-reported native speakers of American English and of participants whose responses to one of the two control trials was more than 2 sd away from the group mean (Exps. 1–3) or whose responses to one of the two control trials was wrong (Exp. 4). Table 3 shows how many participants were excluded in each experiment, demographic information for the remaining participants, and the number of data points that entered into the analyses.

	exclusion criterion		**remaining participants**		data
	language	fillers	ages (mean age)	f/m/nb/dnd	points
Exp. 1 (QUD)	1	10	18-81 (41.1)	36/32/0/1	621
Exp. 2 (asking whether)	2	4	22-74 (38.7)	45/27/1/1	666
Exp. 3 (direct dissent)	2	7	18-77 (39.5)	44/25/1/1	639
Exp. 4 (yes, but)	4	4	19-67 (38.5)	43/27/2/0	648

Table 3: Information from Exps. 1-4 about the number of participants whose data was excluded based on their self-declared language (variety) and the fillers, about the remaining participants, and about the number of data points that entered into the analysis.

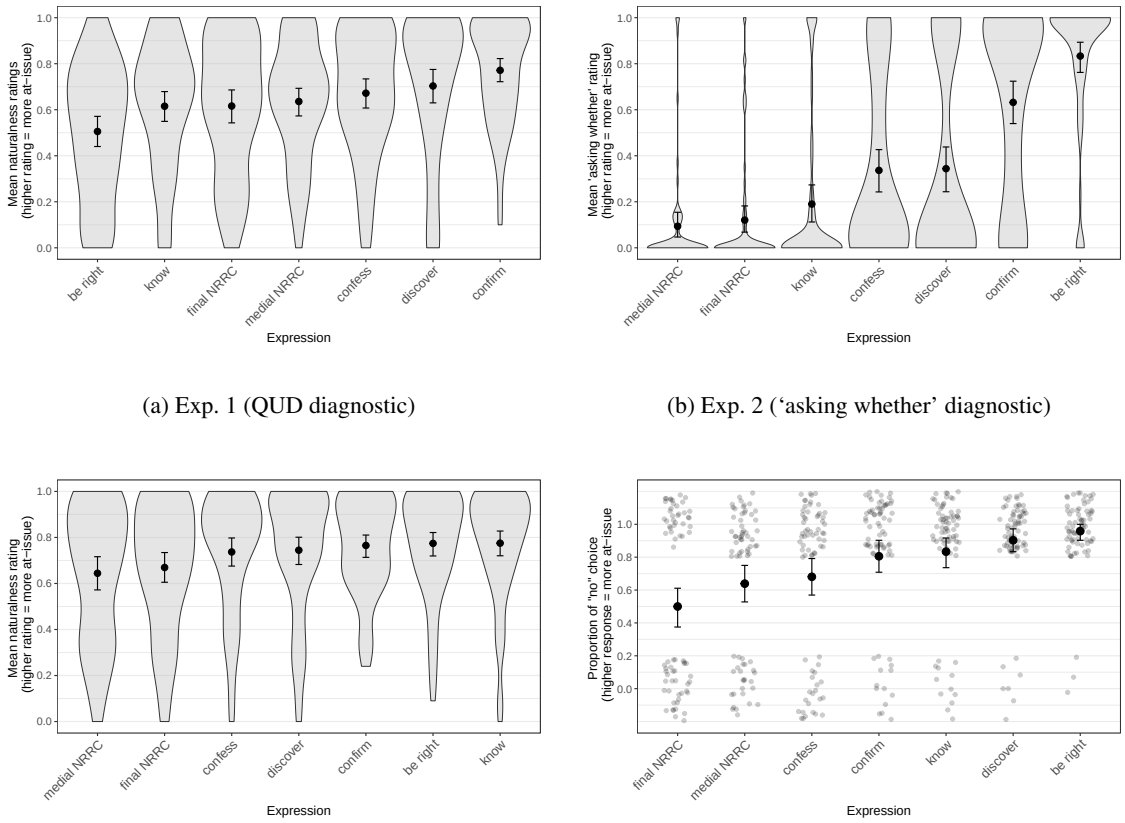
2.2 Results

Fig. 3 plots the results of the four experiments by the expression that is associated with the seven target contents: panel (a) shows the mean naturalness ratings in Exp. 1 (QUD diagnostic), panel (b) mean 'asking whether' ratings in Exp. 2 ('asking whether' diagnostic), panel (c) mean naturalness ratings in Exp. 3 (direct-dissent diagnostic) and panel (d) the proportion of 'no' choices in Exp. 4 ('yes, but' diagnostic).

JT: at this point, the reader expects to read about the results of the comparison. section 1 alluded to one way in which the results of the diagnostics would be compared, namely by whether they show the differences that previous research found, that is, between medial and final NRRCs, and between the five CCs of the clause-embedding predicates. but that's not what the reader gets now; rather, they read about "range of by-content means", which is totally surprising to the reader. my proposal would be to use this range description to help the reader digest the four panels but to not make such a big deal about it as putting it in a subsubsection suggests. also, i am not comfortable with already calling this a result that suggests anything: a diagnostic could have a very small range, but if the means for the individual contents were very small, the diagnostic might still differentiate between the contents.

2.2.1 Range of by-content means

We observe that the results of the four experiments differ in the range of the (mean or proportion of) ratings, that is, the difference between the largest and smallest means. The range is largest in Exp. 2 ('asking whether' diagnostic), at .74 (.1 to .83) and smallest in Exp. 3 (direct-dissent diagnostic), at .13 (.64 to .78). The results of Exp. 1 (QUD diagnostic, with a range of .27 (.51 to .77) and Exp. 4 ('yes, but' diagnostic), with a range of .46 (.5 to .96), fall in between. These results suggest that the four diagnostics, as implemented here, differ in how much they differentiate between the seven



(a) Exp. 1 (QUD diagnostic)

(b) Exp. 2 ('asking whether' diagnostic)

(c) Exp. 3 (direct-dissent diagnostic)

(d) Exp. 4 ('yes, but' diagnostic)

Figure 3: Results of Exps. 1–4. Panels (a)–(c) show the mean responses by expression for (a) Exp. 1 (QUD diagnostic), (b) Exp. 2 ('asking whether' diagnostic), and (c) Exp. 3 (direct-dissent diagnostic); panel (d) shows the proportion of 'no' choices by expression in Exp. 4 ('yes, but' diagnostic). Error bars indicate 95% bootstrapped confidence intervals. Violin plots in panels (a)-(c) show the kernel probability density of individual participants' ratings. Gray dots in panel (d) represent individual participant responses ('no' vs. 'yes', jittered vertically and horizontally for legibility).

contents investigated, with the 'asking whether' diagnostic showing the most differentiation and the direct-dissent and the QUD diagnostic showing the least.**JT: what i mean with "differentiation" is whether the diagnostic suggests that two contents differ in at-issueness, not whether the means are spread out widely across the range. those are related, but different things.**

2.2.2 Rank order and Spearman rank correlations

JT: as i've pointed out above, the intro says that the contents were chosen because previous work found that they differ on some diagnostic and we'll look at whether they are also differentiated in these four diagnostics. so that would be something the reader is looking for now. this discussion here isn't about that, it is about differences between the diagnostics but not in a way that the reader will appreciate, i think; perhaps the ordering business can be mentioned after the main result is presented?

We also find that the four experiments differ in the relative ratings assigned to the seven contents, with only limited overlap. Across all experiments, *confirm* and *discover* consistently received higher ratings (at least numerically) than *confess*, which in turn received higher ratings than medial and final NRRCs. For all other pairs of expressions, however, the ordering was not consistent. In particular, the ranking of *be right* relative to most other contents is inconsistent across experiments (e.g., compared to appositive NRRCs): The embedded content of *be right* received the lowest ratings in Exp. 1, but was among the most at-issue in the other three experiments. The embedded content of *know* was rated (numerically) lower than that of *confirm* in Exps. 1 and 2, but higher in Exps. 3 and 4.

This difference between the results of the experiments is quantified in the Spearman rank correlations shown in Table 4.[6]

	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Exp. 1 (QUD diagnostic)		.11	-.29	-.18
Exp. 2 ('asking whether' diagnostic)			.64	.79
Exp. 3 (direct-dissent diagnostic)				.79

Table 4: Spearman rank correlations between the results of Exps. 1-4.

The rank correlations are particularly low for Exp. 1 compared to the other three experiments, at least partly because of the low relative ranking of *be right*. These results suggests that the four diagnostics as implemented in Exps. 1-4 interact differently with the seven contents investigated.

2.2.3 Pairwise differences between expressions

Finally, the experiments differ in whether they reproduce distinctions between contents reported in prior literature.**JT: i think this should be first: first describe the differences for sentence medial and final nrrcs, and the five CCs, then present the stats** Fig. 4 presents the results of post-hoc pairwise comparisons of the estimated means/proportions for each content, using the 'emmeans' package (Lenth 2023) in R (R Core Team 2016). The input to the pairwise comparisons were mixed-effects beta regression models (Exps. 1-3) or a mixed-effects logistic regression model (Exp. 4). All models were fit using the 'brms' package (Bürkner 2017) using weakly informative priors. The models predicted the ratings[7] from a fixed effect of expression (with treatment coding and 'be right' as reference level) and included random by-participant and by-item intercepts. The output of the pairwise comparison were 95% highest density intervals (HDIs) of estimated marginal mean differences between each of the expressions. We assume that two contents differ if their HDI does not include 0.[8]

Recall that Syrett & Koev 2015, using a variant of the direct-dissent diagnostic, found that sentence-medial NRRCs are more not-at-issue than sentence-final ones. In contrast, as shown in Figs. 3 and 4, no such difference is observed in Exps. 1-3; and in Exp. 4 ('yes, but' diagnostic), we

[6] The Spearman rank correlation coefficient, a value between -1 and 1, is a nonparametric measure of rank correlation: the higher the absolute value of the coefficient, the more the relation between the two variables can be described using a monotonic function. If the coefficient is positive, the value of one variable tends to increase with an increase in the other. In the case of our experiments, a coefficient of 1 for two experiments would mean that there is a perfectly monotone increasing relation between the mean ratings of the seven contents in the two experiments: for any two contents c1 and c2, if c1 ranks below c2 in one experiment (that is, the mean rating of c1 is lower than that of c2), then that ranking is preserved in the other experiment.

[7] To model the ratings in Exps. 1-3 using a beta regression, the ratings were first transformed from the interval [0,1] to the interval (0,1) using the method proposed in Smithson & Verkuilen 2006.

[8] The full model outputs are available in the folder `results+analysis/exps1-4/` in the repository linked in the Data accessibility statement.

find the opposite: sentence-final NRRCs are rated less at-issue than sentence-medial ones. Thus, none of the diagnostics as implemented in Exps. 1-4 replicate Syrett & Koev's finding.

Recall also that Degen & Tonhauser 2025, using the 'asking whether' diagnostic, observed the following at-issueness differences among the content of the complement of clause-embedding predicates: *know* < *discover* < *confess* < *confirm* < *be right* (where the embedded content of *know* is least at-issue, and that of *be right* is most at-issue). As shown in Figs. 3 and 4, Exp. 2 ('asking whether') largely replicates this pattern, except that *confess* and *discover* do not differ.

In Exp. 1 (QUD diagnostic), the embedded content of *confirm* is more at-issue than that of *confess, know,*, but the only other difference observed (among clause-embedding predicates) is that the embedded content of *be right* is less at-issue than those of the other predicates – the direction of this difference is the opposite from that observed in prior literature and the other experiments. In Exp. 3, no differences between the embedded contents of clause-embedding predicates are found. Finally, in Exp. 4, the content of the complement of *be right* is more at-issue than those of *confirm, confess,* and *know*, that of *confirm* is more at-issue than that of *confess*, and that of *discover* is more at-issue than that of *confess*. These results suggest that the diagnostics, as implemented in Exps. 1-4, differ in whether they indicate differences in at-issueness between the contents of the complements of the five clause-embedding predicates included in the experiments.

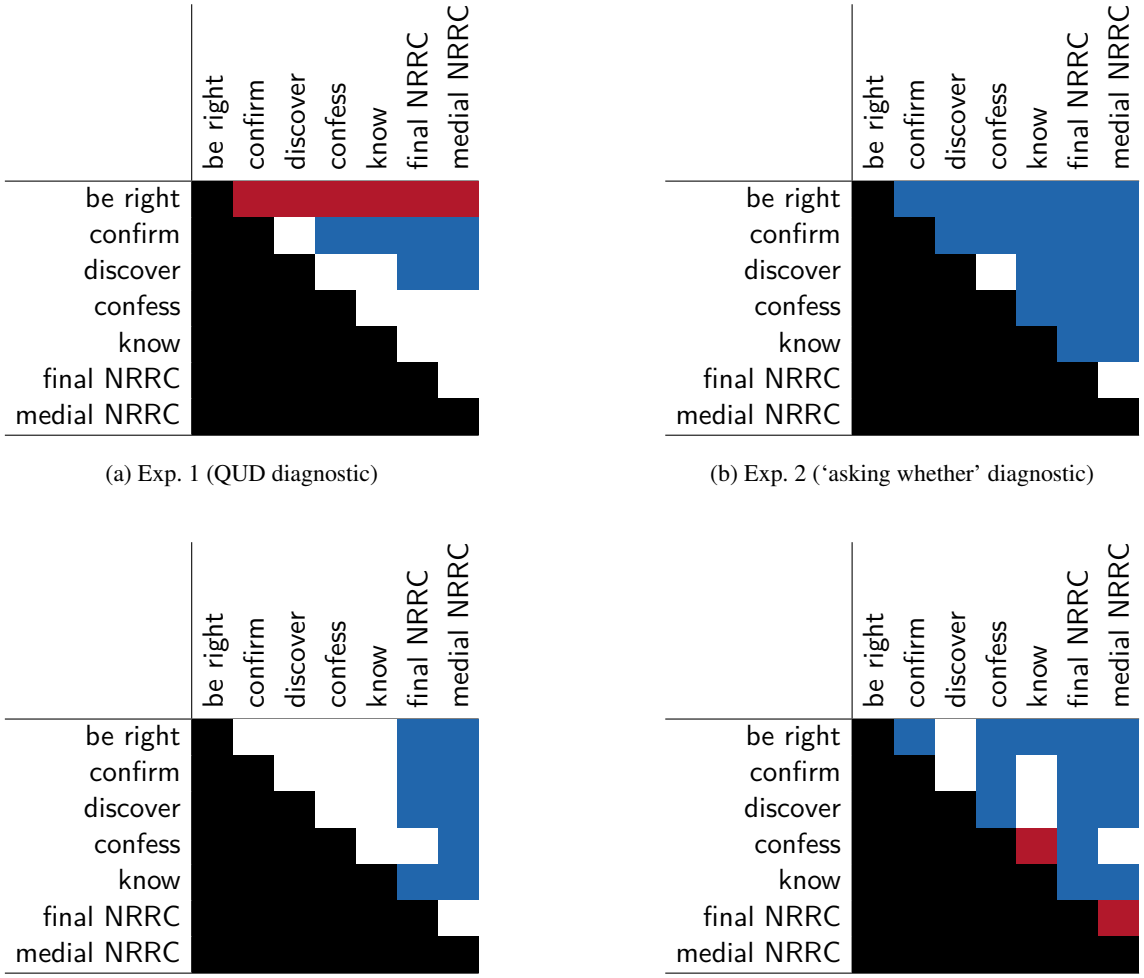
These results also lend further support to the above reported finding that the four diagnostics, as implemented here, differ in how much they differentiate between the seven contents investigated. In particular, while none of the experiments found differences between medial and final NRRCs, the results of Exp. 2 ('asking whether' diagnostic) distinguished between most of the contents of the complements of the five clause-embedding predicates, whereas the results of Exp. 3 (direct-dissent diagnostic) distinguished between the least of these contents. In line with the range of means/proportions reported above, the pattern of differences between contents suggests that the 'asking whether' diagnostic showed the most differentiation, while the direct-dissent and QUD diagnostics showed the least.

2.3 Discussion

Exps. 1-4 were designed to compare the results of four different diagnostics of at-issueness that have been used in prior literature. The results of the experiments suggest that the diagnostics, in the particular way in which they were implemented in the experiments, differ on several dimensions.

i. They differ in the extent to which they differentiate between the seven contents investigated, as evidenced by differences in the range of mean ratings and the numbers of reliable by-content differences. The 'asking whether' diagnostic (Exp. 2) showed the most differentiation and the direct-dissent diagnostic (Exp. 3) showed the least.
ii. They differ in the relative order of the seven contents. For example. the relative ranking of *be right* and other contents (e.g., appositive NRRCs) differs across Exps. 1–4.
iii. They differ with respect to where they find by-content differences: For instance, Exps. 1–3 did not find the positional effect reported in Syrett & Koev 2015 for appositive NRRCs under a forced choice direct-dissent diagnostic. In our study, Exp. 4, using the 'yes but' test, a forced-choice direct-response diagnostic, found an effect opposite to theirs: appositive NRRCs were more at-issue in medial position, compared to final ones.

This section offers discussion of these findings, particularly addressing the question why the 'asking whether' exhibits the most differentiation among the diagnostics, why *be right* in particular ranks high under all diagnostics except for the QUD diagnostic, and how the differences between diagnostics we found bear on whether they target distinct underlying notions of at-issueness.



(a) Exp. 1 (QUD diagnostic)

(b) Exp. 2 ('asking whether' diagnostic)

(c) Exp. 3 (direct-dissent diagnostic)

(d) Exp. 4 ('yes, but' diagnostic)

Figure 4: Pairwise differences between expressions, ordered from top to bottom and left to right by increasing mean in Exp. 2 ('asking whether' diagnostic). A white cell means that the 95% HDI of the pair of the row expression and the column expression includes 0, a red cell means that the 95% HDI does not include 0 and that the coefficient is positive (the row expression received a higher rating than the column expression), and a blue cell means that the 95% HDI does not include 0 and the coefficient is negative (the row expression received a lower rating than the column expression).

2.3.1 Why does the 'asking whether' diagnostic show greater differentiation?

One of the most noticeable differences between the results of the experiments is that Exp. 2 ('asking whether') showed the greatest differentiation between contents, whereas the other three experiments exhibited a smaller range of means, and fewer statistically reliable differences between investigated contents. Among the first 4 experiments, Exp. 2 was the only one that came close to differentiating fine-grained by-predicate differences as reported in Degen & Tonhauser (2025).

To explain this difference, appeal to the divide emphasized in prior work between question-based diagnostics (QUD, 'asking whether') and assertion-based diagnostics (direct-dissent, 'yes, but'):**JT: is this an imperative ('appeal to...') or is something missing here?** Question-based tests are typically taken to probe whether a proposition is at-issue relative to a QUD introduced

in the discourse,**JT: this is very far away from how Tonhauser et al 2018 characterized the asking-whether diagnostic** whereas assertion-based diagnostics are assumed to rely on different assumptions: Snider (2017a; b; 2018) suggests that they reflect the anaphoric availability of propositional content to be antecedents for response particles, which may be independent of at-issueness, while others suggest that these diagnostics target a distinct notion of at-issueness relative to the proposal made by an assertion (Koev 2018; Faller 2019; Korotkova 2020). However, this distinction could not explain why the 'asking whether' diagnostic behaves differently from the QUD diagnostic, since both are considered question-based.

Instead, we must look to something that is present in the 'asking whether' test,**JT: diagnostic?** but not the others to explain the greater by-content differentiation. Two crucial differences set apart the 'asking whether' test from the other three: First, it is the only diagnostic where the target content itself occurs inside a polar question, whereas the other diagnostics test content embedded in declarative assertions. Second, it is the only diagnostic in which participants are asked directly about the intentions of the speaker, whereas the other ones collect measures that are more indirectly tied to semantic interpretation, such as acceptability ratings and continuation choices. Either of these distinctive characteristics could account for the greater differentiation. This yields two hypotheses, which we tested in Exps. 5–6, presented in Section 3.

2.3.2 Do ranking differences suggest different underlying notions?

JT: this subsubsection is about the odd rating for be right in Exp 1 but the subsubsection title suggests otherwise

The diagnostics, as implemented here, yield different rank orders regarding which contents are judged more or less at-issue. One might initially take this to suggest that different diagnostics in fact track different theoretical notions of at-issueness. However, the most striking ranking difference, between the low ranking of the complement of *be right* in Exp. 1 (QUD diagnostic) and its high ranking in the other three experiments, does not appear to reflect a conceptual difference. Instead, it arises from how the QUD diagnostic interacts with additional discourse requirements of *be right*. As shown in panel (a) of Fig. 3, participants gave relatively low naturalness ratings to responses like that in (10a), suggesting that they did not view the response as fitting the question.

(10) a. Exp. 1 (QUD diagnostic) with *be right*
Nora: *What did Lucy break?*
Leo: *Danny is right that she broke the plate.*

b. Exp. 2 ('asking whether' diagnostic)
Nora: *Is Danny right that she broke the plate?*

c. Exp. 3 ('direct dissent' diagnostic)
Nora: *Danny is right that she broke the plate.*
Leo: *No, she didn't break the plate.*

d. Exp. 4 ('yes, but' diagnostic)
Nora: *Danny is right that she broke the plate.*
Leo: *Yes, but she didn't break the plate.*
Yes, and she didn't break the plate.
No, she didn't break the plate.

We hypothesize that this is because *be right* in (10) presupposes that Danny has previously committed to the proposition that "Lucy broke the plate" (Abusch 2010; Anand & Hacquard 2014). In the three diagnostics in (10b–d), no previous discourse context is given, so this presupposition can be accommodated. In contrast, the preceding question in the QUD-diagnostic (10a) conflicts with the presupposition, making it difficult to accommodate. Specifically, we hypothesize that *be*

right signals that the question "whether Lucy broke the plate" is salient in the preceding discourse. This allows us to understand the ill-formedness based on QUD-based discourse structure, because this presupposed question is a subquestion (in the sense of Roberts 1996) of the question in (10a) "What did Lucy break", since every complete answer to the latter entails an answer to the former. The progression from the presupposed subquestion question to the explicitly given superquestion violates Roberts' constraint on QUD stacks (p. 6:15, (10giii)), which allows only the reverse order (from superquestions to subquestions).

As a result, low QUD-match ratings for (10a) are predicted, not because the embedded clause fails to be at-issue, but because the utterance presupposes an incoherent discourse context.[9] This highlights that diagnostics may interact with contextual requirements independently of at-issueness, and such interactions must be taken into account when interpreting the results.

Turning to other ranking differences, Fig. 4 shows that *know* was less at-issue than *confess* in Exp. 2 ('asking whether'), but more at-issue than *confess* in Exp. 4 ('yes but'). We can, again, ask whether this distinction can be attributed to different underlying theoretical notions of at-issueness, such as between question-based and assertion-based conceptions of at-issueness. This interpretation is not ruled out by our data – Exps. 1 (QUD) and 3 (direct dissent) did not detect a difference between *know* and *confess*, but higher-powered studies might reveal that the question-based diagnostics on the one hand and the assertion-based diagnostics on the other come to the same results. However, it is not clear how the presumed theoretical distinction would predict why *confess* and *know* in particular should diverge in this way. jtthis feels weak and speculative to me, i don't think we should include this. i also don't think we should discuss these rank order differences at all, given that several of them are not significant, as per Fig 4

Ultimately, the by-content rank differences in Exps. 1–4 do not provide conclusive evidence about whether there are distinct underlying notions of at-issueness, while the different behaviors found for *be right* show that diagnostics may interact differently with the contextual requirements of the target contents in ways that may be independent from at-issueness.

2.3.3 Positional effects for appositive NRRCs?

As mentioned above, none of our experiments replicated the positional effect reported in Syrett & Koev's Exp. 2, which found that sentence-final appositive NRRCs were judged more at-issue than sentence-medial ones under a variant of the direct-dissent diagnostic.

used a forced-choice task in which participants selected a dissenting response targeting either the matrix clause or the appositive content, illustrated in (11).

(11) Syrett & Koev 2015, p. 17

A: My friend Sophie, a classical violinist, performed a piece by Mozart.
B1: No, she's not. (target: NRRC)
B2: No, she didn't. (target: main clause)

and second, because the direct dissent diagnostic as implemented in Syrett + koev uses response options between directly dissenting with the target content (i.e., the NRRC content in (11)) versus with the main clause content. That different from the version in xue and onea, and implemented here, where all response options reject the target content, and choices are between using direct dissent (*no*), indirect dissent (*yes, but*), and assent (*yes, and*). In our Exp. 3, which also used a

[9] When *be right* is excluded, the Spearman rank correlations are:

	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Exp. 1 (QUD diagnostic)		.77	-.09	-.31
Exp. 2 ('asking whether' diagnostic)			.66	.66
Exp. 3 ('direct dissent' diagnostic)				.77

version of the 'direct dissent' diagnostic, no difference between medial and final appositive NRRCs was found, and Exp. 4 ('yes but') found the opposite effect to Syrett & Koev: here, medial NRRCs were more at-issue than final ones. For comparison, consider again how the diagnostics were implemented in Exps. 3 and 4:

(3) Exp. 3 ('direct dissent' diagnostic)

Nora: *Greg, who bought a new car, is envied by his neighbor.*
Leo: *No, that's not true, he didn't buy a new car.*

(4) Exp. 4 ('yes, but' diagnostic)

Nora: *Greg, who bought a new car, is envied by his neighbor.*
Leo: *Yes, but he didn't buy a new car.* /
Yes, and he didn't buy a new car. /
No, he didn't buy a new car.

Although the diagnostics implemented in Exps. 3 and 4 were dissent/assent-based tasks, one notable difference between our experiments and the Syrett & Koev study, aside from the stimuli, is the response task. Exp. 3 (direct-dissent) elicited naturalness ratings for a direct-dissent response rejecting the target content (3), rather than asking whether it would be preferred over rejecting the main clause content. Exp. 4 ('yes but') collected forced-choice continuation judgments for direct-dissent responses, like Syrett & Koev, but the alternative choices were different: For a response contradicting the target content, participants chose between expressing dissent (*no*), expressing assent and contrast (*yes, but*), or expressing assent and no contrast (*yes, and*).

Taken together, these findings suggest that positional effects for appositive NRRCs may not be robust across diagnostics, and even small changes in the response task used to operationalize a diagnostic, such as alternative choices in a forced-choice continuation study, might reverse the observed pattern, and different diagnostics vary in whether they detect any positional difference for appositive NRRCs at all. This suggests that previously reported effects may reflect task- or diagnostic-specific artifacts rather than stable properties of NRRC interpretation. A natural follow-up would be to test whether applying their particular forced-choice schema to our stimuli would reproduce their effect.

JT: i agree with you on this point but i don't think it warrants its own subsubsection, and it could be dramatically shorter, like a single paragraph, especially if (9) is in the intro

2.3.4 Interim conclusion

JT: i don't think we need interim conclusions: the results are really not that hard to keep in mind and this interrupts the flow from the point about the asking-whether diagnostic to motivating experiments 5-6

The findings from Exps. 1–4 show that the four diagnostics are not interchangeable: they differ in how strongly they differentiate among the tested contents, in the relative ranking between them, and in how they interact with discourse requirements imposed by particular expressions.

One source of diagnostic differences arises from interactions between diagnostics and discourse requirements of particular expressions. As discussed in Section 2.3.2, the low Q+A match ratings for the complement of *be right* in Exp. 1 do not reflect low at-issueness, but rather a conflict between the QUD-diagnostic and a presupposition that the embedded proposition was already under discussion. Such presuppositional requirements can depress ratings independently of at-issueness and therefore mask the intended diagnostic signal.

A second source of diagnostic differences may stem from differences in response tasks, even within dissent-based diagnostics. As discussed in Section 2.3.3, we did not replicate Syrett & Koev's positional effect for appositive NRRCs, and Exp. 4, using the assent/dissent-based yes-but

diagnostic (Exp. 4), showed the a positional effect in the opposite direction. This suggests that even small response task differences may make a difference, such as different alternative choices in a dissent-based forced-choice continuation task.

Further, our data do not provide clear evidence for or against this possibility that diagnostic differences reflect different underlying notions of at-issueness. As discussed in Section 2.3.2, Some by-predicate ranking differences emerged, but these were did in line with the distinctions suggested in the literature and remain inconclusive.

Finally, the largest source of diagnostic differences among Exps. 1–4 concerns the 'asking whether' diagnostic (Exp. 2), which showed strikingly greater differentiation among contents than the other three diagnostics. As discussed in Section 2.3.1, it yielded a wider range of mean ratings and a greater number of reliable contrasts than the other three diagnostics, and it was the only test that approached the fine-grained predicate-level differences reported in Degen & Tonhauser (2025). This raised the question of whether the increased sensitivity is due to interrogative embedding or to the specific response task.

The next section addresses this question directly, reporting on two further experiments, which compare the asking-whether diagnostic to a diagnostic that also embeds the target content in a polar question but uses naturalness ratings for a direct response to assess at-issueness. Both experiments yielded similarly fine-grained distinctions among contents – a finding we take as evidence that the greater variation observed in Exp. 2 is driven primarily by interrogative embedding rather than the response task.

3 Experiments 5 and 6

Exps. 5 and 6 investigate whether the greater differentiation observed in Exp. 2 ('asking whether') was due to (i) embedding target contents in polar questions or (ii) the response task, testing two hypotheses:

i. **Question-embedding hypothesis:** Differentiation between contents (in terms of range of means and significant differences**JT: as i've mentioned above, these don't always go together; i'd focus on the 2nd**) is greater when contents are embedded in a polar question than in a declarative assertion.**JT: 'polar question' is a sentence type; 'declarative assertion' is both sentence type and speech act**

ii. **Response-task hypothesis:** Differentiation is greater when participants are asked directly what the utterance is about, compared to tasks such as acceptability judgments or forced-choice continuations.

To test these hypotheses, we compared two diagnostics that both embed the target content in a polar question but differ in the response task. Exp. 5 used the 'asking whether' diagnostic as in Exp. 2 (12a) and Exp. 6 uses a 'direct response' diagnostic, shown in (12b), where participants read a dialogue between two named speakers, where the first utters a polar question, which contributed the target content.**JT: this sentence is too long** Like in the direct-assent/dissent diagnostic, the second speaker utters 'yes' answer and affirms the target content.

(12) Implementation of the diagnostics in Exps. 5–6

a. Exp. 5 ('asking whether' diagnostic)
Nora: *Is Tom right that Lucy broke the plate?*
Question to participants: Is Nora asking whether Lucy broke the plate?

b. Exp. 6 (direct-response diagnostic)
Nora: *Is Tom right that Lucy broke the plate?*
Leo: *Yes, she didn't break the plate.*
Question to participants: How natural is Leo's response to Nora's question?

Both experiments measured at-issueness for**JT: of?** the contents of the complements of the 20 clause-embedding predicates from Tonhauser et al. 2018 and Degen & Tonhauser 2025, listed in (13).**JT: if they need to be repeated (which i don't think they do), then they should be ordered alphabetically**

(13) 20 clause-embedding predicates from Tonhauser et al. 2018; Degen & Tonhauser 2025:
be annoyed, know, pretend, inform, see, hear, discover, acknowledge, think, admit, announce, reveal, confess, demonstrate, suggest, prove, establish, say, confirm, be right

Both experiments also measured projection data**JT: measured the projection of?** for the embedded contents**JT: ccs?** of the 20 clause-embedding predicates, which were reported in Hofmann et al. 2024. Here we focus on the at-issueness data, which have not yet been reported.

If the greater differentiation between contents found in Exp. 2 (compared to Exps. 1, 3, and 4) was due to presenting the target content in a polar question, we expect that results from Exps. 5 and 6 both show a similar differentiation between contents – namely comparable range of means and comparable sensitivity to differences between contents.**JT: what if only one of these came out? and how does one measure comparable range of means? we have no stats to back this up. i would omit the range considerations. also, the phrasing "comparable sensitivity to differences between contents" presupposes that the contents differ and we're just trying to find a diagnostic that shows these differences. but i don't think we should be conveying that presupposition** In contrast, if the 'asking whether' diagnostic again shows greater differentiation between contents than the direct-response diagnostic, this would support the response-task hypothesis, suggesting that the task, rather the question embedding, drives the effect.

3.1 Methods

3.1.1 Participants

We recruited 300 participants on Amazon's Mechanical Turk platform for Exp. 5 and 250 participants on Prolific for Exp. 6.[10] Participants recruited on Mechanical Turk had U.S. IP addresses and at least 99% of previous HITs approved. Participants recruited on Prolific had registered as U.S.-born native speakers of English residing in the USA and had an approval rate of at least 99%. Table 5 summarizes the age and gender distributions of the recruited participants.

	recruited	ages (mean age)	f/m/nb/dnd
Exp. 5 (asking whether)	300	19-74 (38.2)	–/–/–/–
Exp. 6 (direct assent)	250	18-58 (25.5)	201/43/6/0

Table 5: Information about the participants recruited in Exps. 5-6 (f = female, m = male, nb = nonbinary, dnd = did not disclose; gender information was not collected in Exp. 5).

3.1.2 Materials and procedure

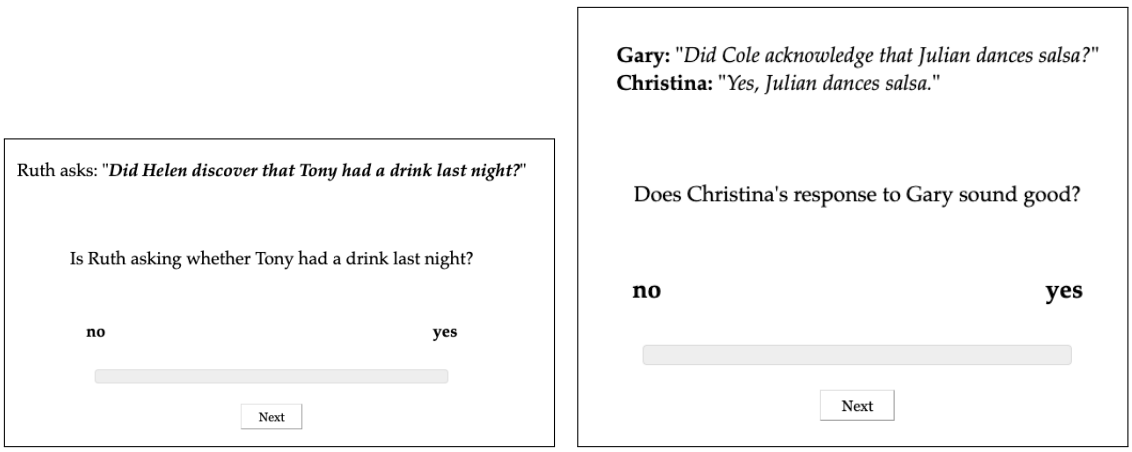
Each of the 20 clause-embedding predicates in (13) was combined with one of 20 embedded-clause items**JT: they were referred to differently above** (listed in Supplement B). The two experiments also included the same six control stimuli.**JT: the introduction of the target stimuli should be completed first: how many target stimuli in each experiment? what was their shape in each experiment? etc** The contents of these control stimuli (details in Supplement C) were expected to be at-issue and were used to assess participants' attention. In both experiments, each participant's

[10] Exp. 5 was run in August 2019 and Exp. 6 in August 2021.

set of items was generated by randomly combining each of the 20 clause-embedding predicates in (13) with a unique item. Participants saw a total of 26 stimuli: one target stimulus for each of the 20 clause-embedding predicates and the same 6 control trials.[11] Trial order was randomized for each participant.

Participants were asked to imagine that they are at a party and that, when walking into the kitchen, they overhear somebody say something to somebody else.**JT: this might be easier if you refer to Fig 5**

As shown in Fig. 5, the response task differed between experiments. In Exp. 5 ('asking whether' diagnostic, panel (a)), participants judged whether the question was about the target content, using a slider marked 'no' on one end (coded as 0) and 'yes' on the other (coded as 1). In Exp. 6 (direct-response diagnostic, panel (b)), participants rated whether the response to the first speaker sounds good, on a slider marked 'no' (coded as 0) on one end and 'yes' on the other (coded as 1). Across both experiments, the responses were coded so that 1 meant that the content to be diagnosed was rated as at-issue and 0 as not-at-issue.**JT: would it be better to talk about higher ratings meaning higher not-at-issueness? how was this done in exps 1-4?**



(a) Exp. 5: 'asking whether' diagnostic

(b) Exp. 6: direct response diagnostic

Figure 5: Sample trials in (a) Exp. 5 and (b) Exp. 6.

JT: something needs to be said about the projection ratings, the two blocks, the order of the two blocks, etc.

After completing the experiment, participants filled out a short optional demographic survey. To encourage truthful responses, participants were told that they would be paid no matter what answers they gave in the survey.

3.1.3 Data exclusion

We excluded the data of participants who did not self-identify as native speakers of American English, of participants whose responses to the projection or at-issueness controls were more than 2 sd away from the group mean, and of participants who always selected roughly the same point on the response scale for the target stimuli.**JT: did we do this latter one in exps 1-4 as well?** To identify such participants, we first identified participants whose mean variance on the target stimuli was more than 2 sd below the group mean variance and then manually inspecting their response

[11] Each participant saw their set of 26 stimuli twice, once in the projection block and once in the at-issueness block. Block order was randomized. As mentioned above, we focus here on the at-issueness data.

patterns. Due to a programming error, 5 participants took Exp. 5 more than once. Since we were not able to identify which submission was their first submission, the data of these participants was also excluded. Table 6 shows how many participants were excluded in each experiment, the properties of the remaining participants, and the number of data points that entered into the analyses.

	exclusion criterion			remaining participants			data
	language	controls	variance	ages (mean age)	f/m/nb/dnd	total	points
Exp. 5 (asking whether)	7	35	0	21-74 (39.2)	–/–/–/–	242	6292
Exp. 6 (direct assent)	5	24	1	18-58 (24.9)	187/28/5/0	220	5720

Table 6: Information from Exps. 5-6 about the number of participants whose data was excluded based on their self-declared language and language variety ('language'), the controls, and the variance of their responses, about the remaining participants, and about the number of at-issueness data points that entered into the analysis.

3.2 Results

Fig. 6 plots the results of the two experiments by embedding predicate: panel (a) shows the mean 'asking whether' ratings in Exp. 5 ('asking whether' diagnostic) and panel (b) shows the mean acceptability ratings in Exp. 6 (direct-response diagnostic).

JT: can the two plots in Fig 6 go side by side?

3.2.1 Range of by-content means

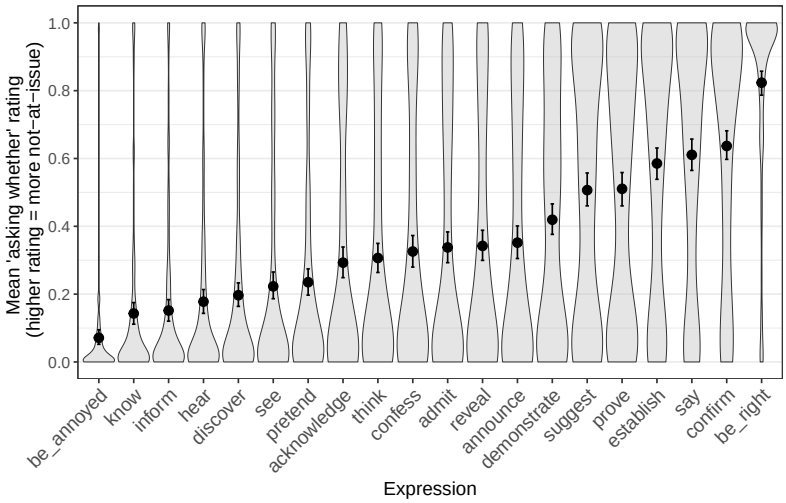
We observe that the results of the four **JT: two?** experiments show a similar range of the mean ratings (again quantified as the difference between the largest and smallest by-content means). The range in Exp. 5 ('asking whether' diagnostic) is .75 (.07 to .82) and in Exp. 6 (direct-response diagnostic), it is .73 (.09 to .82). While results of Exps. 1–4 showed clear variation in the range of the by-content means, the same cannot be said about the results of Exps. 5 and 6.**JT: to me, this is not the most exciting point and it doesn't warrant its own subsubsection. if anything, i'd include it somewhere after the main result.**

3.2.2 Rank order and Spearman rank correlations

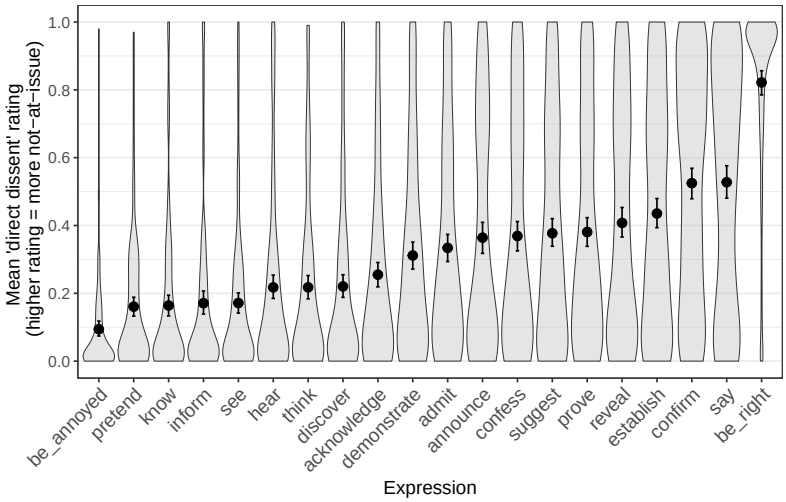
We also find a high degree of consistency in the relative ratings across the 20 contents: predicates whose embedded content is rated relatively at-issue in Exp. 5 are also rated relatively at-issue in Exp. 6, and vice-versa.This correspondence is reflected in a Spearman rank correlation of .93 between the results of Exps. 5 and 6, suggesting a particularly strong rank correlation for the by-content means. This result suggests that the two diagnostics as implemented in Exps. 5 and 6 interact similarly with the 20 target expressions investigated, when those are presented in a question embedding.**JT: same comment as above**

3.2.3 Pairwise differences between expressions

Fig. 7 presents the results of post-hoc pairwise comparisons of the estimated means for each content. As in Section 2, these were done using the 'emmeans' package (Lenth 2023) in R (R Core Team 2016), based on mixed-effects beta regression models, that were fit using the 'brms' package (Bürkner 2017) using weakly informative priors. The models predicted ratings from a fixed effect of expression (with treatment coding and 'be right' as the reference level).**JT: need that footnote again about converting [0,1] to (0,1)**



(a) Exp. 5 ('asking whether' diagnostic)

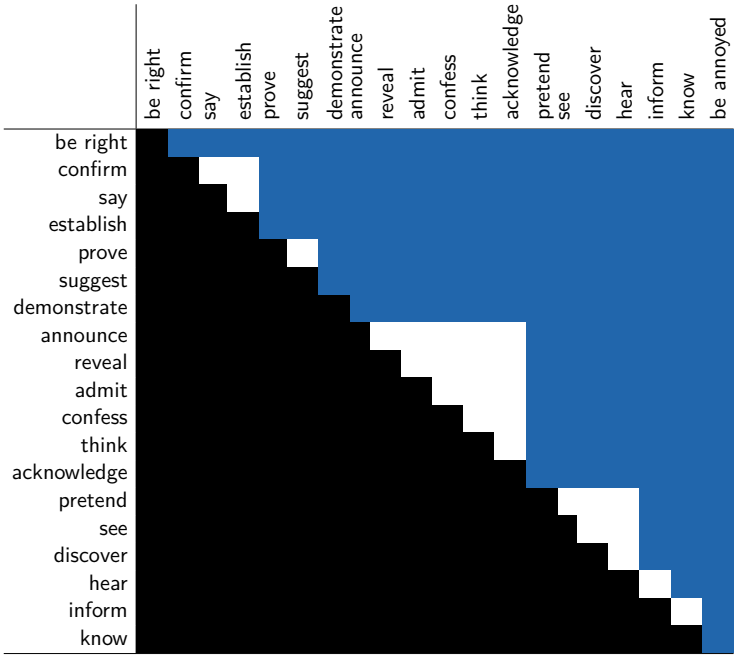


(b) Exp. 6 ('direct assent' diagnostic)

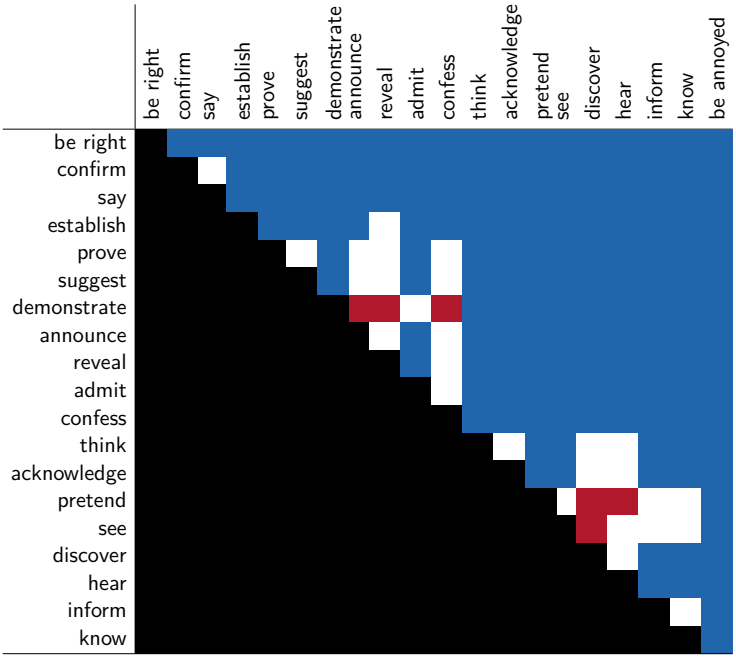
Figure 6: Results of Exps. 5–6. The panels show the mean ratings by expression for (a) Exp. 5 (asking whether diagnostic) and (b) Exp. 2 (direct assent diagnostic). Error bars indicate 95% bootstrapped confidence intervals. Violin plots show the kernel probability density of individual participants' ratings.

Overall, the two experiments show broadly similar results: both reveal reliable pairwise distinctions among many of the expressions and **JT: yield comparable rank orderings – isn't this a repetition?**. The ordering among the clause-embedding predicates found in Degen & Tonhauser 2025*know* < *discover* < *confess* < *confirm* < *be right* (which were largely replicated in Exp. 2) is found in both Exps. 5 and 6.**JT: this too?, this isn't about the results shown in Fig 7 but about the rank orderings?**

There are some differences between the two experiments: In Exp. 5, the complement predicate *demonstrate* is more at-issue than the complement of a range of predicates, including *announce,*



(a) Exp. 5 ('asking whether' diagnostic)



(b) Exp. 6 ('direct assent' diagnostic)

Figure 7: Pairwise differences between expressions, ordered from by increasing mean in Exp. 5 ('asking whether'). White cells indicate that the 95% HDI of the difference includes 0. Red cells indicate a positive difference (the row expression received a higher rating than the column expression), and blue cells indicate a negative difference.

reveal, and *confess*, but in Exp. 6, the embedded content of *demonstrate* is comes out as less at-issue than the content of those three predicates.

Other differences between the two experiments include that in Exp. 5 *prove* and *suggest* are more at-issue than *reveal* and *confess*, *think* and *acknowledge* are more at-issue than *discover* and *hear*, *pretend* and *see* are more at-issue than *inform* and *know*, where Exp. 6 finds no differences. Conversely, in Exp. 6 *confirm* and *say* are more at-issue than *establish*, *pretend* is less at-issue than *discover* and *hear*, *see* is less at-issue than *discover*, where Exp. 5 finds no difference.

There are, however, some minor differences between the two experiments.**JT: didn't the text above also point out differences?** In Exp. 5, the complement of *demonstrate* is rated as more at-issue than those of *announce*, *reveal*, and *confess*, whereas in Exp. 6 it is rated as less at-issue than the complement of those same predicates.

In the results of Exp. 5, several differences also come out as reliable,**JT: what does 'reliable' mean?** where no difference is observed in Exp. 6: (i) *prove* and *suggest* are more at-issue than *reveal*; (ii) *confess*, *think*, and *acknowledge* are more at-issue than *discover* and *hear*; and (iii) *pretend* and *see* are more at-issue than *inform* and *know*. Conversely, Exp. 6 reveals differences not found in Exp. 5, including that: (i) *confirm* and *say* are more at-issue than *establish*; (ii) *pretend* is less at-issue than *discover* and *hear*; and (iii) *see* is less at-issue than *discover*.

JT: I don't think all the differences need to be spelled out here; the reader can look for themselves. we should hit the highlights.

Taken together, these results suggest that while the diagnostics as implemented in Exps. 5 and 6, may differ slightly in which distinctions they detect, their overall patterns of at-issueness judgments**JT: no at-issueness judgments were given** are highly similar.

3.3 Discussion

Our results suggest that the 'asking whether' and direct response diagnostic, as implemented in Exps. 5 and 6 yielded highly similar results for the 20 contents investigated: both showed comparable ranges of by-content means, a particularly strong Spearman rank correlation ($r_s = .93$), and fine-grained differentiation among the 20 contents tested. This supports the idea that their shared feature of presenting the target content embedded in a question is what drives the high by-content differentiation observed in Exps. 2, 5, and 6 compared to Exps. 1, 3, and 4.

Nonetheless, small differences remain, for example in which by-content contrasts reached significance and in the precise rank ordering of contents. Since both tests used identical embedding contexts and both reflect a QUD-based conception of at-issueness, these differences may stem from the distinct response tasks used: judging what a question is about versus (Exp. 5) evaluating the naturalness of a response (Exp. 6). This suggests that differences in response task affect at-issueness judgments in subtle ways.

JT: this is not a discussion but a summary of the results, plus very brief discussion statements. perhaps fold them into the results subsection, which one could rename 'results and discussion'?

4 General discussion

Different diagnostics of at-issueness yield different results. Some of the differences between the diagnostics as implemented here appear to be due to presenting the target expressions embedded in **JT: polar** questions, some of them have to do with how the diagnostics interact differently with the discourse requirements imposed by the tested expressions,**JT: is this about "be right"? if so, that's a super tiny point compared the point about interrogative vs. declarative – should it really be mentioned here on the same level?** and others may have to do with response task differences,**JT: which of our results support this?** or different underlying notions of at-issueness.**JT: very little has been said about notions of at-issueness; i don't see this as a point on par with the**

interrogative vs. declarative one In our investigation, the factor leading to the most striking differences was whether or not the tested target content is embedded in a question, so the speech act in which it appears**JT: do contents appear in speech acts? also, the particular speech act was not mentioned yet, unless "question" for you is both a sentence type and a speech act?** needs to be taken into consideration when considering what at-issueness is**JT: this is a very big claim, not one that should appear in this summary of results. perhaps you mean "when considering how at-issueness is diagnosed?"** and how it is diagnosed.**JT: oh, i guess you really meant "what at-issueness is".** In the following, we discuss in some more detail the questions of why question embeddings matter for diagnosing at-issueness (Section 4.1); what are factors that may affect the diagnostic differences, and what this can tell us about different underlying notions of at-issueness (Section 4.2); then we discuss methodological implications of our findings (Section 4.3).

4.1 Speech act, at-issueness, and projection

We found that the speech act in which the tested content appears makes a difference:**JT: did we? or is that a hypothesis for which we found support? i also think that the reader needs to be guided more towards the declarative/assertion vs. polar interrogative/question (or info-seeking?) connection** presenting the target expression in a question leads to greater by-content differentiation than presenting it in an assertion. As discussed briefly in Section 2.3.1, this contrast may seem to reflect the contrast highlighted in previous literature between question-based and assertion-based diagnostics for at-issueness.**JT: isn't the contrast described in prior literature as qud- vs proposal-based?** Recall that question-based diagnostics (such as the QUD-diagnostic and 'asking whether' test) have been taken to assess whether a proposition is at-issue relative to a question under discussion (Amaral et al. 2007; Simons et al. 2010; Tonhauser 2012; Tonhauser et al. 2018). In contrast, assertion-based diagnostics (such as the direct-dissent and 'yes, but' tests) have been suggested to target a different notion of at-issueness, tied to the speech act of assertion (Koev 2018; Faller 2019; Korotkova 2020).**JT: i disagree with this characterization of the literature. i think you're going too far away from the literature too quickly** Here, the assumption is that only at-issue content can contribute an assertive proposal that may be accepted or rejected, while not-at-issue content is presupposed or automatically added to the common ground (Potts 2005; Murray 2014; AnderBois et al. 2015).

Our generalization, however, is not about whether the target content can be understood as at-issue relative to a QUD versus an assertion made by the utterance.**JT: I don't understand this sentence** Rather, it concerns the speech act in which the content is presented.**JT: again, are contents presented in speech acts?** When target contents were presented in questions, as in the 'asking whether' diagnostic (Exps. 2, 5) and the direct-response diagnostic used in Exp. 6, we observed high by-content differentiation. In contrast, when the same contents were presented in declarative assertions, as in the QUD diagnostic (Exps. 1), and the assertion-based diagnostics (Exps. 3, 4), the resulting ratings showed less differentiation. This pattern suggests that at-issueness diagnostics which embed target content in questions are suited to reveal more differentiated by-content differences than those embedding content in assertions.

This finding can be understood if we assume that participants are better able to distinguish at-issue from not-at-issue content when these two contents have more distinct pragmatic roles,**JT: 'pragmatic roles' is not a term i am familiar with** and questions provide an environment where that is the case. We hypothesize the relevant contrast**JT: between what?** lies in how at-issue and not-at-issue content relate to speaker commitments: In assertions, but not in questions, the speaker commits to the truth of the at-issue proposition. Not-at-issue content, in contrast, projects**JT: not from positive assertions** and thus contributes to speaker commitment regardless of speech act (see Potts 2005; Abusch 2010; Simons et al. 2010; Abrusán 2011; Tonhauser et al. 2018; Degen & Tonhauser 2025). As a result, at-issue and not-at-issue content are pragmatically more distinct in

questions,**JT: "pragmatically more distinct" is very vague** making it easier for participants to distinguish the two levels of meaning.**JT: 'level' implies that at-issue and not-at-issue content live on distinct dimensions?** Embedding target content in questions therefore allows for detection of more fine-grained differences in at-issueness associated with the conventional meaning of particular expressions, whereas assertions introduce a confounding factor: the speaker commits to both levels of content, reducing differentiation in at-issueness judgments.**JT: what's missing here is something on why embedding in questions differentiates better between different not-at-issue contents; so far, this has all been about differentiating at-issue and not-at-issue content (which our experiments did not investigate)**

For concreteness, (14) illustrates how the at-issue content of a polar question introduces a set of question alternatives and partitions the context set without committing the speaker to any particular alternative (Groenendijk & Stokhof 1984; Ginzburg 1996; Roberts 1996). In contrast, the not-at-issue content expressed by the appositive NRRC (*Greg bought a car*) is assumed to be true across all question alternatives, and thus throughout the entire context set, giving rise to projection (Abusch 2010; Simons et al. 2010; Abrusán 2011; Tonhauser et al. 2018).

(14) $[\![$*Is Greg, who bought a car, envied by his neighbor?*$]\!] =$
{Greg, who bought a car, is envied by his neighbor,
Greg, who bought a car, is not envied by his neighbor}

Because the not-at-issue proposition projects, it is treated as part of the speaker's commitments, while the at-issue proposition is not. We suggest that this sharp epistemic contrast between the two contents may make it easier for listeners to distinguish between levels of meaning. Questions therefore provide a particularly a good environment for detecting fine-grained differences in at-issueness associated with the conventional meaning of particular expressions.**JT: this repeats what was said above, almost verbatim**

JT: the paragraph around 12 is about a sentence-medial NRRC, which is maximally different in terms of at-issueness and speaker commitment from the at-issue content, which is at-issue and the speaker is not committed. you suggested above that speaker commitment is relevant. so when we have content that is not as clearly not-at-issue as sentence-medial NRRCs, like the CC of "discover" or of "reveal", why does embedding in questions help bring out differences between them? i don't see how the example in 12 helps explain this, or something is missing

JT: does this imply that embedding under negation should also not work so well (commitment to negated at-issue content) but embedding under "perhaps" (or another epistemic modal) should? don't we have experiments like Exp 6 but with embedding under "perhaps" and with embedding under negation?

This reasoning helps explain why the two question-based diagnostics yielded highly similar results. Both the asking whether diagnostic used in Exps. 2 and 5 and the direct-response diagnostic used in Exp. 6 embed the target content *p* in a polar question, as in (15), where *p* is the proposition that *Tony had a drink last night* introduced by the complement of *discover*.

(15) *Did Helen confirm that Tony had a drink last night?*

In the 'asking whether' diagnostic, ratings of how much participants take the question to be about *whether p* (in this case: about whether *Tony had a drink last night*) are taken to reflect whether *p* is delineates the question partition introduced by the at-issue content of (15), and thus how readily *p* is interpreted as at-issue. Similarly, Exp. 6 assessed how naturally a direct response to (15) with *yes p* or *no, not p* is taken as an answer to the question, that is, how naturally they are construed as selecting among the question alternatives. Both diagnostics therefore probe how central *p* is to the question raised by the utterance.**JT: i don't understand how this paragraph around (13) relates**

to the hypothesis about question embedding; is it supposed to further support it? it seems to merely explain the two diagnostics?

Thus, if certain constructions or expressions bias their associated content toward being interpreted as at-issue or not-at-issue**JT: this sounds very categorical, which is not consistent with our results**, interrogatives provide a clear window onto these by-content differences.**JT: this does not follow from what has been said above** For instance, since the complement of *confirm* received higher asking-whether ratings in Exp. 2 than the content of appositive NRRCs, we can say that the complement of *confirm* is more at-issue than the content of appositive NRRCs (or more likely to be interpreted as at-issue under a categorical conception of at-issueness). An important question for research on at-issueness is why certain expressions lead to their associated content being interpreted as more or less at-issue, which lexical properties can affect this and how (e.g., Abrusán 2011; Anand & Hacquard 2014; Schlenker 2021; Anand & Korotkova 2024; Bade 2024; Bade et al. 2024).

In contrast, when the same contents are presented in declarative assertions, shown (16) and (17), as in the QUD diagnostic (Exp. 1) and the assertion-based diagnostics (Exps. 3 and 4), the utterance signals that speaker is committed to the main clause at-issue content.

(16) *Greg, who bought a car, is envied by his neighbor.*
  a. → *Greg bought a car.* (target content)
  b. → *Greg is envied by his neighbor.* (main clause content)

(17) *Helen confirmed that Tony had a drink last night.*
  a. → *Tony had a drink last night.* (target content)
  b. → *Helen confirmed that Tony had a drink last night.* (main clause content)

At the same time, not-at-issue content, like that expressed by appositive NRRCs or certain embedded propositions, projects**JT: there is no entailment-canceling operator, at least in 14, and therefore no projection. and i'm not comfortable saying that the CC projects in 15 from under "confirmed" – this is very far away from how people talk about projection; speaker commitment is fine, of course** and thus also becomes part of the speaker's commitments. In assertions, the speaker is thus committed to both the main at-issue and the not-at-issue content.**JT: i'm not following at all. are you saying that the speaker of 15 is committed to the CC? and the same with all the other clause-embedding predicates?** As a result, at-issue and not-at-issue content are pragmatically more similar**JT: "pragmatically more similar" is too vague** in assertions, and we suggest that this makes it harder for participants to distinguish the two levels of meaning. This assumption could helps explain why these diagnostics yielded less by-content differentiation than the question-based diagnostics: they might be overall less sensitive to the difference between at-issue and not-at-issue content, essentially creating a ceiling effect.

To our knowledge, no previous literature has proposed that the speech act in which target content is presented can itself be a major driver of differentiation. Here, we saw that questions more transparently reveal how different contents can be interpreted as at-issue or not-at-issue, and suggested that this may be because the two contents differ in their commitment status. Assertions, in contrast, commit the speaker both to the at-issue assertion as well as the projected not-at-issue content, thereby obscuring the distinction. Consequently, diagnostics that embed the target content in questions provide a particularly sensitive test of at-issueness.

4.2 Diagnostic differences and notions of at-issueness

Our study addressed two related questions about at-issueness diagnostics: First, we asked to what extent different diagnostics yield consistent results when applied to the same linguistic stimuli. Second, by examining the convergence and divergence across diagnostics, we aimed to assess

whether these differences reflect distinct theoretical notions of at-issueness or instead arise from methodological differences in how a single underlying phenomenon is operationalized.**JT: as far as i'm concerned, the second question has not yet been addressed. i think it may be better to make the paper just about the first one, and bring the second one up in the General Discussion only. of course, the business about some diagnostics being about q-at-issueness and some about p-at-issueness needs to stay in section 1**

Across Exps. 1–4, we found clear evidence that the four diagnostics implemented there are not interchangeable: they interact differently with the seven tested contents, and the relative rank orderings of content means do not align uniformly across experiments. Aside from the robust ordering *confess* $\leq$ *discover*, no pairwise difference between contents goes in the same direction across all diagnostics. At the same time, the most striking differences between diagnostics appear to be explainable by factors other than distinct underlying notions of at-issueness.**JT: this paragraph just repeats the results but it's not clear why – what's the question being addressed?**

First,**JT: what are we counting? rhetorical structure is unclear** as discussed in detail in Section 4.1, the speech act in which the target content is presented plays a major role. Diagnostics that embed the target content in questions consistently yielded greater differentiation than those that embed it in declarative assertions. This finding, however, is orthogonal to the question whether there are distinct underlying notion of at-issueness:**JT: so why start with that?** it is compatible both with the view that all diagnostics target a single underlying notion and with the view that there are fundamentally different concepts of what the main point of an utterance is given a question or QUD versus given an assertive proposal.**JT: i think it's worthwhile repeating the q- and p-based definitions of at-issueness here from section 1 and starting from them** We argued here that the effect arises because questions and assertions differ in how at-issue and not-at-issue content interact with speaker commitment. Questions sharpen this contrast, whereas assertions collapse it, thereby reducing differentiation. The influence of speech act thus offers an explanation for a substantial portion of the observed divergence between diagnostics.**JT: what is this paragraph doing? it seems to be repeating what was said in 4.1 while also saying that this doesn't mean that there's different underlying notions of at-issueness. is this the best way to get into the discussion?**

Second, the diagnostics differ in how they interact with contextual requirements imposed by particular expressions. A clear example is the behavior of *be right* under the QUD diagnostic, where low QUD-match ratings do not plausibly reflect not-at-issueness, but instead arise from a pragmatic incompatibility: *be right* presupposes a discourse structure that conflicts with the assumptions made by the diagnostic (see Section 2.3.2). A related concern is raised by Snider (2017a; 2018), who emphasized that the direct-dissent and 'yes but' diagnostic involve propositional anaphora in the interpretation in the response particles *yes/no* (as does the direct-response diagnostic in Exp. 6). Ratings under these diagnostics may therefore reflect constraints on anaphoric availability that are independent of at-issueness. Although our data show no clear split between diagnostics that involve response particles (Exps. 3, 4, and 6) and those that do not, the broader point remains: different diagnostics impose different requirements on discourse structure, common ground, and anaphoric accessibility, all of which can independently affect acceptability judgments. This supports the conclusion that some differences between diagnostics arise from interactions with expression-specific contextual requirements rather than from differences in at-issueness per se.**JT: i see, i guess you're going through the differences in our results between the diagnostics and discussing, for each one, whether that means that there are different underlying notions of at-issueness? if yes, this rhetorical structure needs to be made clearer. i didn't get that from the above "...we aimed to assess whether these differences reflect distinct theoretical notions of at-issueness or instead arise from methodological differences in how a single underlying phenomenon is operationalized" because that was in the past tense (at least "aimed") and to me that sounded like we had already done it (which we hadn't; see my comment above).**

Third, some small differences between diagnostics can be attributed to response-task effects. This is evident in the subtle differences between Exps. 5 and 6 (see section 3.3), which differed only in response format, and in our failure to replicate Syrett & Koev's 2015 reported positional effect for appositive NRRCs (see section 2.3.3). While they found sentence-final appositives to be more at-issue than sentence-medial ones under a version of the direct-dissent diagnostic, our assent/dissent-based *yes, but* diagnostic in Exp. 4 showed a difference in the opposite direction, while the other experiments found no difference between them (Exp.1–3). Taken together, these results provide no evidence for a stable notion of at-issueness under which sentence-final appositives are systematically more at-issue than medial ones (or the other way round), and they highlight how even small task differences can affect outcomes.**JT: does this paragraph argue that response task effects don't point to different underlying notions of at-issueness? or does it go off track and argue that sentence-medial vs -final NRRCs may not differ in at-issueness?**

While many differences between our experimental results find plausible explanations in factors that are independent of whether there are separate underlying notions of at-issueness, our data showed no obvious distinction between QUD-at-issueness and proposal at-issueness.**JT: this was said many times already and should not need to be repeated here; if anything, the reader could be reminded at the beginning of the subsection, when the two notions are repeated.** This could be due to artifacts of the items or the instructions that may have led to obscuring an underlying distinctions. At the same time, among most pairs from Exps. 1–4, the Spearman rank correlation was at least moderate (when excluding *be right*, see footnote 9), suggesting that the diagnostics, as implemented there, are at least partially sensitive to a shared underlying property.

Our data does not provide conclusive evidence about whether or not the diagnostics track fundamentally different underlying notions of at-issueness. However, the fact that many of the differences find explanations outside of this question, seems more compatible with assuming that the diagnostics offer different, partially overlapping windows onto a single discourse phenomenon, that interacts with multiple discourse factors (speaker commitments, speech acts, contextual requirements, and the given alternatives) in complex and sometimes subtle ways, so that even small differences in task design can make a difference.

4.3 Methodological implications

JT: this is too short for its own subsection. it's also one of our major results, so it should be at the beginning of the General Discussion section.

Our findings also have methodological implications for future empirical research on at-issueness. The diverging results suggest that the diagnostics cannot be applied interchangeably, and theoretical claims made on their basis should be relativized to the diagnostic used (see also Snider 2017b; a; 2018; Koev 2018; Korotkova 2020). For instance, if a study finds that a particular construction is more at-issue than another using one diagnostic, this conclusion may not hold if a different diagnostic is employed. Therefore, empirical investigations should continue path of using multiple diagnostics**JT: see above** and future research on at-issueness should carefully consider that results may be specific to the speech act context,**JT: what is a speech act context?** contextual requirements of the expressions used, or the particular response task design.**JT: response task?** In particular, our findings underscore the importance of considering the speech act context in which target contents are presented. As we have shown, embedding target contents in questions versus assertions can greatly impact the results of at-issueness diagnostics.**JT: this is too repetitive.**

5 Conclusion

JT: for me, this is too long and rehashes too many of the not-so-major results that we found empirical support for. i would just write one paragraph that repeats the main research question, the main result, the main implication and the main big question that remains (at-issueness definitions)

In a series of six experiments, this study investigated how different diagnostics for at-issueness behave when applied to the same linguistic stimuli, using two sets of stimuli,**JT: is it critical that we used two sets of stimuli?** and five**JT: right** different diagnostics: the QUD diagnostic, the 'asking whether' diagnostic, the direct-dissent diagnostic, the 'yes, but' diagnostic, and the 'direct response' diagnostic. Our findings provide experimental confirmation for claims that there are empirical differences between at-issueness diagnostics (Snider 2017b; a; 2018; Koev 2018; Faller 2019; Korotkova 2020). The diagnostics yielded different results, with some diagnostics showing greater differentiation among contents than others.

We identified several factors contributing to these differences, including the speech act in which the target content is presented, interactions with contextual requirements imposed by specific expressions, and properties of the response task. Although our results suggest that the choice and implementation a diagnostic matter for empirical generalizations about at-issueness, the theoretical divide between QUD-based and assertion-based diagnostics assumed in previous literature does not appear to be the primary source of divergence. Instead, the speech act context plays a central role: embedding the target content in questions leads to greater differentiation among contents than embedding it in assertions. This finding highlights the need to consider the speech act context when interpreting results from at-issueness diagnostics.

Additional factors, such as contextual requirements of the particular expressions used, and the response task design can also influence at-issueness judgments in complex ways. The diversity of discourse factors influencing at-issueness judgments highlights the need for careful consideration of diagnostic methods in empirical research on at-issueness. Future work should take into account interactions between diagnostics, speech acts, contextual requirements, and response tasks when selecting diagnostics and items for their studies. Having identified some factors that can influence results, there remains an important question for future work: what the results of different diagnostics might reveal about whether or not they reflect a shared underlying notion of at-issueness.

Abbreviations (if applicable)

NRRC = non-restrictive relative clause, QUD = question under discussion **JT: these should be explained in the text and then this bit can be deleted**

Data accessibility statement

The experiments, data and R code for generating the figures and analyses of the experiments reported in this paper are available at https://anonymous.4open.science/r/at-issueness-diagnostics-232C/.

Ethics and consent

All experiments were conducted with approval from the ethics review committee of [university name redacted for review].

References

Abbott, Barbara. 2000. Presuppositions as nonassertions. *Journal of Pragmatics* 32(10). 1419–1437. https://doi.org/https://doi.org/10.1016/S0378-2166(99)00108-3. Publisher: Elsevier

Abrusán, Márta. 2011. Predicting the presuppositions of soft triggers. *Linguistics and Philosophy* 34(6). 491–535. https://doi.org/10.1007/s10988-012-9108-y

Abusch, Dorit. 2010. Presupposition triggering from alternatives. *Journal of Semantics* 27(1). 37–80.

Amaral, Patricia & Roberts, Craige & Smith, E Allyn. 2007. Review of the logic of conventional implicatures by Chris Potts. *Linguistics and Philosophy* 30. 707–749. Publisher: Springer.

Anand, Pranav & Hacquard, Valentine. 2014. Factivity, belief and discourse. In Crni, Luka & Sauerland, Uli (eds.), *The Art and Craft of Semantics: A Festschrift for Irene Heim*, 69–90. MIT Working Papers in Linguistics. https://semanticsarchive.net/Archive/jZiNmM4N/.

Anand, Pranav & Korotkova, Natasha. 2024. Facts, intentions, questions: English come-to-knowpredicates in deliberative environments. In *Proceedings of the Amsterdam Colloquium*. 15–21. https://platform.openjournals.nl/PAC/article/view/21764.

AnderBois, Scott & Brasoveanu, Adrian & Henderson, Robert. 2010. Crossing the appositive/at-issue meaning boundary. In *Semantics and linguistic theory*, vol. 20. 328–346. http://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/2551.

AnderBois, Scott & Brasoveanu, Adrian & Henderson, Robert. 2015. At-issue proposals and appositive impositions in discourse. *Journal of Semantics* 32(1). 93–138. https://doi.org/10.1093/jos/fft014

Bade, Nadine. 2024. New data on the 'triggering problem'for presuppositions. In *Semantics and Linguistic Theory*. 197–212. https://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/34.010.

Bade, Nadine & Schlenker, Philippe & Chemla, Emmanuel. 2024. Word learning tasks as a window into the triggering problem for presuppositions. *Natural Language Semantics* 32(4). 473–503. https://doi.org/10.1007/s11050-024-09224-5. https://link.springer.com/10.1007/s11050-024-09224-5

Barnes, Kathryn & Ebert, Cornelia. 2023. The information status of iconic enrichments: modelling gradient at-issueness. *Theoretical Linguistics* 49(3-4). 167–223. https://doi.org/10.1515/tl-2023-2009. https://www.degruyter.com/document/doi/10.1515/tl-2023-2009/html

Beaver, David & Clark, Brady. 2008. *Sense and Sensitivity: How Focus Determines Meaning*. Oxford: Wiley-Blackwell. https://doi.org/10.1002/9781444304176

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). 1–28. https://doi.org/10.18637/jss.v080.i01

Büring, Daniel. 2003. On D-Trees, Beans, And B-Accents. *Linguistics and Philosophy* 26(5). 511–545. https://doi.org/10.1023/A:1025887707652. https://link.springer.com/10.1023/A:1025887707652

Chen, Yuqiu. 2024. *Presuppositions at the Semantics-Pragmatics Interface*: Georg-August-Universität Göttingen Doctoral dissertation. https://ediss.uni-goettingen.de/handle/11858/15164.

Clark, Herbert H. & Schaefer, Edward F. 1989. Contributing to discourse. *Cognitive science* 13(2). 259–294. https://www.sciencedirect.com/science/article/pii/0364021389900086.

Cummins, Chris & Amaral, Patricia & Katsos, Napoleon. 2013. Backgrounding and accomodation of presuppositions: An experimental approach. In *Proceedings of Sinn und Bedeutung*, vol. 17. 201–218.

Degen, Judith & Tonhauser, Judith. 2025. Projection inferences: On the relation between prior beliefs, at-issueness, and lexical meaning. Manuscript under review.

Destruel, Emilie & Onea, Edgar & Velleman, Daniel & Bumford, Dylan & Beaver, David. 2015. A cross-linguistic study of the non-at-issueness of exhaustive inferences. In Schwarz, Floran (ed.), *Experimental approaches to presupposition*, 135–156. Springer. https://doi.org/10.1007/978-3-319-07980-6

Esipova, Maria. 2019. *Composition and projection in speech and gesture*. New York, NY: New York University Doctoral dissertation. https://www.researchgate.net/publication/371702326_Composition_and_projection_in_speech_and_gesture.

Esipova, Maria. 2021. On not-at-issueness in pictures. *Glossa: a journal of general linguistics* 6(1). https://www.glossa-journal.org/articles/10.5334/gjgl.1314/?utm_source=TrendMD&utm_medium=cpc&utm_campaign=Glossa%253A_a_journal_of_general_linguistics_TrendMD_0.

Faller, Martina. 2006. Evidentiality below and above speech acts. https://semanticsarchive.net/Archive/GZiZjBhO/info.txt.

Faller, Martina T. 2002. *Semantics and pragmatics of evidentials in Cuzco Quechua (Peru)*: Stanford University Doctoral dissertation. https://www.proquest.com/dissertations-theses/semantics-pragmatics-evidentials-cuzco-quechua/docview/305523331/se-2?accountid=10957.

Faller, Martina T. 2019. The discourse commitments of illocutionary reportatives. *Semantics and Pragmatics* 12. 8:1–53. https://doi.org/10.3765/sp.12.8. https://semprag.org/index.php/sp/article/view/sp.12.8

Farkas, Donka & Bruce, Kim. 2010. On reacting to assertions and polar questions. *Journal of Semantics* 27(1). 81–118. https://doi.org/10.1093/jos/ffp010. https://scholar.archive.org/work/hymiwpqf5fdazop52liak5zuoa/access/wayback/https://people.ucsc.edu/~farkas/papers/assertion_question.pdf

Ginzburg, Jonathan. 1995. Resolving Questions I, II. *Linguistics & Philosophy* 18. 459–527, 567–609.

Ginzburg, Jonathan. 1996. Interrogatives: Questions, facts and dialogue. *The handbook of contemporary semantic theory* 5(18). 359–423. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=595cd12adcf5e27f900c47778695412e89711481. Publisher: Citeseer.

Groenendijk, Jeroen Antonius Gerardus & Stokhof, Martin Johan Bastiaan. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*: Univ. Amsterdam PhD Thesis.

Hofmann, Lisa & de Marneffe, Marie-Catherine & Tonhauser, Judith. 2024. Projection variation: Is the family of sentences really a family? *Sinn und Bedeutung* 28. 422–440.

Horton, Diane & Hirst, Graeme. 1988. Presuppositions as beliefs. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*. https://aclanthology.org/C88-1052.pdf.

Hunter, Julie & Asher, Nicholas. 2016. Shapes of Conversation and At-Issue Content. *Semantics and Linguistic Theory* 1022–1042. https://doi.org/10.3765/salt.v26i0.3946. https://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/26.1022

Jasinskaja, Katja. 2016. Not at issue any more. *Ms. University of Cologne* https://dslc.phil-fak.uni-koeln.de/sites/dslc/katja_files/jasinskaja_any_more.pdf.

Karttunen, Lauri & Peters, Stanley. 1979. Conventional implicature. In Oh, Choon-Kyu & Dinneen, David A. (eds.), *Presuppositions* (Syntax and Semantics Vol.11), 1–56. New York: Academic Press.

Koev, Todor. 2018. Notions of at-issueness. *Language and Linguistics Compass* 12. e12306. https://doi.org/https://doi.org/10.1111/lnc3.12306

Koev, Todor K. 2013. *Apposition and the structure of discourse*. Rutgers The State University of New Jersey, School of Graduate Studies. https://search.proquest.com/openview/3686668834d9802d690c6e574100c8e0/1?pq-origsite=gscholar&cbl=18750.

Korotkova, Natasha. 2020. Evidential meaning and (not-)at-issueness. *Semantics and Pragmatics* 13. article 4.

Lee, Jungmee. 2011. *Evidentiality and its interaction with tense: Evidence from Korean*: The Ohio State University PhD Thesis. https://rave.ohiolink.edu/etdc/view?acc_num=osu1306940284.

Lenth, Russell V. 2023. *emmeans: Estimated marginal means, aka least-squares means*. https://CRAN.R-project.org/package=emmeans. R package version 1.8.8.

Murray, Sarah E. 2010. *Evidentiality and the structure of speech acts*: Rutgers The State University of New Jersey-New Brunswick Doctoral dissertation.

Murray, Sarah E. 2014. Varieties of update. *Semantics and Pragmatics* 7. 2:1–53. https://doi.org/10.3765/sp.7.2. https://semprag.org/index.php/sp/article/view/sp.7.2

Onea, Edgar & Beaver, David. 2009. Hungarian focus is not exhausted. In *Semantics and linguistic theory*. 342–359. http://journals.linguisticsociety.org/proceedings/index.php/SALT/article/download/2524/2272.

Papafragou, Anna. 2006. Epistemic modality and truth conditions. *Lingua* 116(10). 1688–1702. https://www.sciencedirect.com/science/article/pii/S0024384106000805. Publisher: Elsevier.

Potts, Christopher. 2005. *The Logic of Conventional Implicatures*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199273829.001.0001

Potts, Christopher. 2015. Presupposition and Implicature. In Lappin, Shalom & Fox, Chris (eds.), *The Handbook of Contemporary Semantic Theory*, 168–202. Wiley 1st edn. https://doi.org/10.1002/9781118882139.ch6

R Core Team. 2016. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Roberts, Craige. 1996. Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. In Yoon, Jae Hak & Kathol, Andreas (eds.), *Ohio State University Working Papers in Linguistics*, vol. 49. The Ohio State University, Department of Linguistics.

Roberts, Craige. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics & Pragmatics* 5. 1–69.

Schlenker, Philippe. 2021. Triggering presuppositions. *Glossa: a journal of general linguistics* 6(1). https://www.glossa-journal.org/articles/10.5334/gjgl.1352/. Publisher: Open Library of Humanities.

Scontras, Gregory & Tonhauser, Judith. 2025. Projection without presupposition: A model for know. In *Sinn und Bedeutung (SuB) 29*. To appear.

Shanon, Benny. 1976. On the two kinds of presuppositions in natural language. *Foundations of language* 14(2). 247–249. https://www.jstor.org/stable/25170057.

Simons, Mandy & Tonhauser, Judith & Beaver, David & Roberts, Craige. 2010. What projects and why. In *Semantics and linguistic theory*, vol. 20. 309–327.

Smithson, Michael & Verkuilen, Jay. 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* 11. 54–71.
Snider, Todd. 2017a. At-issueness anaphoric availability. *Proceedings of the Linguistic Society of America* 2. 39–1. http://journals.linguisticsociety.org/proceedings/index.php/PLSA/article/view/4089.
Snider, Todd. 2018. Distinguishing at-issueness from anaphoric potential: A case study of appositives. In *West Coast Conference on Formal Linguistics (WCCFL)*, vol. 35. 374–381.
Snider, Todd N. 2017b. *Anaphoric reference to propositions*. Ithaca, NY: Cornell University Doctoral dissertation. https://ecommons.cornell.edu/server/api/core/bitstreams/3794923d-d85f-4c7c-85ac-8106f83152b5/content.
Solstad, Torgrim & Bott, Oliver. 2024. Cataphoric resolution of projective content: The case of occasion verbs. *Semantics and Pragmatics* 17. 11:1–66. https://doi.org/10.3765/sp.17.11. https://semprag.org/index.php/sp/article/view/sp.17.11
Stalnaker, Robert. 1973. Presuppositions. *Journal of Philosophical Logic* 4. 447–57.
Stalnaker, Robert C. 1978. Assertion. In Cole, Peter (ed.), *Pragmatics* (Syntax and Semantics 9), 315–332. Leiden, Netherlands: Brill. https://doi.org/10.1163/9789004368873_013.
Stalnaker, Robert C. 2002. Common ground. *Linguistics and Philosophy* 25. 701–721. https://www.jstor.org/stable/pdf/25001871.pdf.
Syrett, Kristen & Koev, Todor. 2015. Experimental evidence for the truth conditional contribution and shifting information status of appositives. *Journal of Semantics* 32(3). 525–577. https://doi.org/10.1093/jos/ffu007. Publisher: Oxford University Press
Tonhauser, Judith. 2012. Diagnosing (not-) at-issue content. *Proceedings of Semantics of Under-represented Languages of the Americas (SULA)* 6. 239–254.
Tonhauser, Judith & Beaver, David I. & Degen, Judith. 2018. How projective is projective content? Gradience in projectivity and at-issueness. *Journal of Semantics* 35(3). 495–542. https://doi.org/10.1093/jos/ffy007
Xue, Jingyang & Onea, Edgar. 2011. Correlation between presupposition projection and at-issueness: An empirical study. In *Proceedings of the ESSLLI 2011 workshop on projective meaning*. 171–184.

Supplements

A Control stimuli in Exps. 1–4

The examples in (1)-(4) provide the two control stimuli used in each of Exps. 1-4. For the a.-examples, participants were expected to give a 'totally fits' response (Exp. 1), a 'yes' response (Exp. 2), a 'totally natural' response (Exp. 3), and a 'no' response (Exp. 4); for the b.-examples, the opposite response was expected. The numbers after each example identify the mean ratings (Exps. 1-3) or the proportion of 'no' responses (Exp. 4) after excluding participants who did not self-identify as native speakers of American English (but before excluding participants on the basis of these controls), showing that the control stimuli worked as intended.

(1) Control stimuli in Exp. 1 (QUD diagnostic)
   a. Mary: Which course did Ava take?
      John: She took the French course. (.97)
   b. Jennifer: What does Betsy have?
      Robert: She loves dancing salsa. (.07)

(2) Control stimuli in Exp. 2 ('asking whether' diagnostic)

a. Mary: Did Arthur take a French course?
Question to participants: Is Mary asking whether Arthur took a French course? (.96)
b. Robert: Does Betsy have a cat?
Question to participants: Is Robert asking whether Betsy loves apples? (.02)

(3) Control stimuli in Exp. 3 ('direct dissent' diagnostic)

a. Mary: Arthur took a French course.
Lily: No, he took a Spanish course. (.87)
b. Robert: Betsy has a cat.
Maximilian: No, she doesn't like apples. (.05)

(4) Control stimuli in Exp. 4 ('yes, but' diagnostic)

a. Mary: Arthur took a French course.
Lily: Yes, but Lisa loves cats. / Yes, and he didn't take a French course. / No, he didn't take a French course. (.95)
b. Robert: Betsy has a cat.
Maximilian: Yes, but she is good at math. / Yes, and she loves it so much. / No, she doesn't like apples. (0)

B 20 clauses

The contents of the following 20 clauses, which realized the complements of the 20 clause-embedding predicates, were investigated in Exps. 5–6:

i. Mary is pregnant.
ii. Josie went on vacation to France.
iii. Emma studied on Saturday morning.
iv. Olivia sleeps until noon.
v. Sophia got a tattoo.
vi. Mia drank 2 cocktails last night.
vii. Isabella ate a steak on Sunday.
viii. Emily bought a car yesterday.
ix. Grace visited her sister.
x. Zoe calculated the tip.
xi. Danny ate the last cupcake.
xii. Frank got a cat.
xiii. Jackson ran 10 miles.
xiv. Jayden rented a car.
xv. Tony had a drink last night.
xvi. Josh learned to ride a bike yesterday.
xvii. Owen shoveled snow last winter.
xviii. Julian dances salsa.
xix. Jon walks to work.
xx. Charley speaks Spanish.

C Control stimuli in Exps. 5–6

The control stimuli in Exps. 5–6 were the contents of the main clause polar questions in (5). The non-restrictive relative clauses (NRRCs), given in parentheses in (5), were included in Exp. 6, where at-issueness was measured with an assent diagnostic. The control stimuli here consisted of two clauses (like the target stimuli), to allow the relevant speaker to assent with one of two clauses.

(5) Sentences for control stimuli in in question embedding experiments (Exps. 5–6)

a. Do these muffins (, which are really delicious,) have blueberries in them?
b. Does this pizza (, which I just made from scratch,) have mushrooms on it?
c. Was Jack (, who is my long-time neighbor,) playing outside with the kids?
d. Does Ann (, who is a local performer,) dance ballet?
e. Were John's kids (, who are very well-behaved,) in the garage?
f. Does Samantha (, who is really into fashion,) have a new hat?

We expected participants to give low responses on the at-issueness diagnostics for the control stimuli in (5), indicating that the main clause content is at-issue.