

## Testing the tests: A re-evaluation of at-issueness diagnostics across modalities

**Background.** At-issueness is a central notion at the semantics-pragmatics interface. At-issue content is the main claim of an utterance, while not-at-issue content is often considered a peripheral aside. Presuppositions (Karttunen and Peters, 1979), conventional implicatures (e.g., appositives, Potts, 2005), and co-speech gestures (Ebert et al., 2020) have been argued to contribute not-at-issue meaning. Several diagnostics for identifying (not-)at-issue content have been proposed (e.g., Tonhauser, 2012). However, most of these rely on introspective judgments and were originally designed to test spoken expressions. Their generalizability and suitability for assessing the at-issue status of iconic gestures are thus questionable. In fact, varying sensitivity across different diagnostics has already been observed for not-at-issue content in speech (Xu et al., 2025). We therefore argue that the current methodological toolbox is not yet equipped to reliably investigate the at-issue status of co-speech gestures. To test this, we conducted three rating studies using a 0–100 slider scale (0 = min; 100 = max) employing different diagnostics for not-at-issue content, comparing iconic co-speech gestures with appositives, both canonically assumed to contribute not-at-issue meaning. We hypothesize that across all experiments, diagnostics are more sensitive to appositives than iconic gestures. All experiments consisted of 24 experimental items and 24 fillers.

**Experiment 1** tested the mismatch paradigm (Ebert et al., 2020). Items were distributed across four lists in a Latin square design. Sixty native speakers of German participated. Each trial consisted of a short video in which a speaker uttered a sentence containing either a sentence-medial nominal appositive (1a) or a co-speech gesture aligned with the noun (1b) conveying parallel content.

(1) a. Auf diesem Bild ist eine Statue auf einem Sockel, der übrigens rund ist, zu sehen.

b. Auf diesem Bild ist eine Statue auf einem [Sockel] zu sehen. + ROUND

‘In this picture, there is a statue on a [base](, by the way a round one). + ROUND’

Participants then saw a picture that either matched the appositive/gesture content (statue on round base) or mismatched it (rectangular base). This yields a 2×2 design with the factors *MODE* (appositive vs. gesture) and *MATCH* (match vs. mismatch). Participants rated how well the sentence in the video matched the image. The mismatch paradigm serves as an indirect at-issueness diagnostic: picture-matching judgments are assumed to be more strongly affected by at-issue than by not-at-issue content (as per Kroll and Rysling, 2019). In practice, mismatches violating at-issue content should elicit sharply reduced ratings, because the contradiction concerns the utterance’s main point. Mismatches involving not-at-issue content, by contrast, should incur a milder penalty. Assuming that diagnostics are more sensitive for spoken not-at-issue content, we predicted a higher mismatch penalty for appositives than for gestures; a stronger decrease from match to mismatch ratings for appositives than gestures. Appositives exhibited a large mismatch penalty ( $\Delta \approx 38.8$  points), whereas gestures showed a substantially smaller penalty ( $\Delta \approx 7.7$  points). A custom contrast computed from a linear mixed-effects model revealed that the mismatch penalty differed significantly across modes, with a larger penalty for appositives than gestures ( $p < .0001$ ).<sup>1</sup>

**Experiments 2–3** applied the direct denial (DD) diagnostic (Exp. 2; Tonhauser’s (2012) diagnostic #1a) and the indirect denial (ID) diagnostic (Exp. 3; Tonhauser’s (2012) diagnostic #1c). Items were distributed across two lists in a Latin square design. Thirty native speakers of German participated per study. Both experiments manipulated only a single factor: *MODE* (appositive vs. gesture). Experimental items structurally matched those in Exp. 1, except that no picture followed the video. Instead, each trial consisted of again a short video in which a speaker A uttered a sentence containing either a sentence-medial nominal appositive or a co-speech gesture aligned with the noun. Participants then saw a written reply by a second speaker B. In Exp. 2, B’s reply denied the content contributed by the appositive or gesture via a DD response; in Exp. 3, B used an ID response:

<sup>1</sup>Model output (Exp. 1–3): [https://osf.io/yjukp/overview?view\\_only=e7c5b075ba3c4020af232685277973aa](https://osf.io/yjukp/overview?view_only=e7c5b075ba3c4020af232685277973aa).

(2) a. DD: Nein, der Sockel war doch nicht rund. ('No, the base wasn't round.')

b. ID: Ja, aber der Sockel war doch nicht rund. ('Yes, but the base wasn't round.')

Participants rated how well B's reply fit the preceding video utterance. DD should be felicitous only when the denied information is at-issue. ID, by contrast, is predicted to be felicitous only when not-at-issue information is denied information in the *but*-continuation. Linear mixed-effects models revealed main effects of *MODE* for Exp. 2 and 3. In Exp. 2, ratings in the appositive condition were substantially higher than in the gesture condition ( $\Delta \approx 38.6$  points,  $p < .001$ ). In Exp. 3, ratings in the appositive condition were again higher than in the gesture condition ( $\Delta \approx 23.4$  points,  $p < .001$ ).

**Discussion.** The three experiments show that appositive and gestural content diverge across all diagnostics, though how they differ depends on the task. In the mismatch paradigm (Exp. 1), appositives incur a much larger mismatch penalty than gestures, suggesting that appositive material may be interpreted as more at-issue in this task. Appositives also differ from gestures in the denial tasks (Exp. 2–3), but in a way difficult to reconcile with the idealized predictions of the diagnostics. Although DD is designed to target at-issue content, appositives receive unexpectedly high ratings. At the same time, they are also rated highly in ID, following the pattern predicted for not-at-issue content. Thus, appositives simultaneously show profiles associated with at-issue and not-at-issue meaning, indicating that the two denial diagnostics yield conflicting results. Gestural content exhibits a different but likewise noncanonical pattern. It receives low ratings in DD, as predicted for not-at-issue content, but only moderate ratings in ID, falling short of the 'high' acceptability expected under the diagnostic. Overall, the diagnostics do not converge on a coherent pattern, neither for gestures nor for appositives. Rather, they reveal that at-issueness is probed in a highly task-dependent way and that the diagnostics differ markedly in the evidence they yield. Consequently, they should not be treated as interchangeable tools for assessing (not-)at-issueness. Notably, the data for appositives parallel the findings of Kroll and Rysling (2019), who show that speakers adapt their judgments in a goal-oriented manner: when aiming to answer a QUD, for instance, they may disregard the canonical at-issue/not-at-issue divide and consider all information relevant to that goal. A further methodological concern is that denial diagnostics require gestural meaning to be verbalized. Since iconic gestures are often underspecified and do not map neatly onto single lexical expressions, such verbalization risks imposing interpretations that participants may not share. However, the unexpected behavior of appositives in the denial tasks suggests that the challenge is not confined to visual content. Even verbal not-at-issue expressions licensed both DD and ID at high levels, collapsing the very contrast denial diagnostics are designed to capture. This raises the broader question of how reliably such diagnostics track at-issueness in experimental settings. Since they originate from introspective methodology, where linguists assess felicity contrasts in carefully constructed contexts, their experimental implementation introduces additional demands (e.g., explicit ratings, forced verbalization) that may blur the underlying distinctions. Our findings therefore cast doubt on whether denial tests, in their current operationalizations, faithfully reflect the theoretical constructs they target. The mismatch paradigm avoids the need to verbalize the target content and therefore offers a promising alternative, but its strong mismatch penalties for appositives complicate interpretation. One limitation of the present design is the absence of an explicit at-issue control condition. We will therefore replicate all three studies with such a condition for comparison. In sum, the findings highlight the need for developing novel experimental paradigms that are equipped to test for the at-issue status of multimodal as well as spoken content.

**Refs.** Ebert, C., C. Ebert, R. Hörnig 2020. Demonstratives as dimension shifters. *SuB24 Proceedings* · Karttunen, L., S. Peters 1979. Conventional implicature. *Syntax and Semantics* · Kroll, M., A. Rysling 2019. The search for truth: Appositives weigh in. *SALT29 Proceedings* · Potts, C. 2005. *The Logic of Conventional Implicatures* · Tonhauser, J. 2012. Diagnosing (not-)at-issue content. *SULA6 Proceedings* · Xu, C., L. Hofmann, J. Tonhauser 2025. What is at-issueness? XPrag Fest