Comparing the Utility of Scoring Methods for the Hinting Task in A Heterogeneous Clinical Sample

Rory MacAndrew McKemey

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science

In

Clinical Science

Chloe C. Hudson, Chair

Angela Scarpa

Louis Hickman

December 4, 2025

Blacksburg, Virginia

Keywords: hinting task, theory of mind, psychometrics

Comparing the Utility of Scoring Methods for the Hinting Task in A Heterogeneous Clinical Sample

Rory MacAndrew McKemey

Abstract

The Hinting Task is a popular theory of mind measure that has been criticized due to poor psychometric properties. A revised set of scoring criteria has reduced ceilings effects and improved convergent validity of Hinting Task scores in individuals with psychotic-spectrum disorders and matched non-clinical controls. In the current study, we are the first to compare the psychometric properties of the original and revised criteria in a heterogenous clinical sample not characterized by psychotic-spectrum disorders. Given the stringent nature of the revised criteria, we also test the novel hypothesis that participant verbosity may explain differences in performance across scoring criteria. Participants were 173 patients (65% female; 80% non-Hispanic White; *M* age = 34.4, *SD* = 12.9) in a partial hospitalization program. Participants completed the Hinting Task, Reading the Mind in the Eyes Test, Patient Health Questionnaire-9, Prodromal Questionnaire-Brief, and Behavior And Symptom Identification Scale on their first or second day of treatment. Hinting Task performance was scored by independent raters using both criteria. Results demonstrated that revised criteria scores had significantly lower ceiling effects compared to original criteria scores. Convergent validity of Hinting Task performance was partially supported and did not differ between scoring criteria. Revised, but not original scores were impacted by verbosity, such that less verbose participants demonstrated worse performance. In summary, our results suggest that the revised criteria improve one psychometric aspect of the task while simultaneously introducing verbosity as a confounding variable. We recommend controlling for verbosity when implementing the revised criteria in future research.

Comparing the Utility of Scoring Methods for the Hinting Task in A Heterogeneous Clinical Sample

Rory MacAndrew McKemey

General Audience Abstract

Theory of mind is the ability to attribute mental states such as thoughts, feelings, and desires to oneself and others. This skill is important for navigating social situations, and research shows that theory of mind is generally impaired across a variety of mental illnesses. The Hinting Task is a widely used measure of theory of mind. One issue with the Hinting Task is that your average person has a good chance to score perfectly on the task, meaning it may not be good at distinguishing between certain high-scoring individuals. Revised scoring criteria have been previously developed and recommended, which were shown to remove these "ceiling effects" by providing more stringent rules for what is considered a correct response to questions on the task. The primary aim of this study is to compare the original and revised scoring criteria to see which version makes the Hinting Task better at measuring theory of mind in a sample of individuals with various mental illnesses. We also test the novel hypothesis that participant verbosity may explain differences in scores across criteria. Participants were patients receiving mental health treatment in a partial hospital program. Results showed that fewer participants received perfect scores with the revised vs original criteria. Scoring criteria did not affect the Hinting Task's association with relevant variables. Revised, but not original scores were impacted by verbosity, such that less verbose participants demonstrated worse performance. In summary, our results suggest that the revised criteria remove ceiling effects but also make the Hinting Task measure verbosity in addition to theory of mind. We recommend accounting for verbosity when implementing the revised criteria in future research.

Comparing the Utility of Scoring Methods for the Hinting Task in A Heterogeneous Clinical Sample

Rory MacAndrew McKemey

Introduction

It is a widely held view among researchers that social relationships are integral to facilitate human health and fulfillment (e.g., Badcock et al., 2017; Cacioppo et al., 2015). The collection of cognitive processes required to successfully navigate interpersonal interactions is often referred to as social cognition (Frith & Frith, 2008; Green & Leitman, 2008). One fundamental aspect of social cognition—theory of mind—is the ability to accurately ascribe mental states (e.g., thoughts, emotions, intentions, beliefs) to oneself and others. Theory of mind itself may be a multi-faceted construct given that it involves inference of several different mental states (e.g., emotions, beliefs, intentions; Dvash & Shamay-Tsoory, 2014). Inaccurate inference of a variety of mental states is robustly associated with poor social functioning and quality of life (Couture et al., 2006; Dodell-Feder et al., 2014; Thibaudeau et al., 2021; Trojsi et al., 2016).

Theory of mind is a construct of concern not only in the context of social outcomes but also as a risk factor for psychopathology. Emerging research suggests that theory of mind may be an underlying mechanism across all forms of psychopathology (Gur & Gur, 2016), given that it has been found to be impaired nearly ubiquitously across a wide span of psychiatric diagnoses (Cotter et al., 2018). Mentalization-based treatment, an evidence-based treatment for borderline personality disorder, specifically aims to improve theory of mind cognitions and has seen success in treating patients with a wide variety of psychiatric disorders (Bateman & Fonagy, 2013). In summary, individual differences in theory of mind has implications in understanding and treating various psychopathologies.

The Hinting Task was developed in 1995 by Corcoran, Mercer, and Frith to measure theory of mind in people with schizophrenia. Since its inception, it has been widely adopted as a theory of mind measure across many psychiatric populations and non-clinical samples. To administer this task, the experimenter reads aloud a series of ten short stories involving an interaction between two characters. At the end of each story, participants are asked what a character *really* meant when they gave a piece of dialogue (referred to as a "hint"). Participants are scored on their ability to infer the intended meaning of the character's hint. A score of 2 is given if they successfully do so. If they do not correctly infer the intentions of the character, participants are given a second, less subtle "hint" that indicates the speaker's desires. Participants are then asked what one character in the story wants the other character to do. At this point, a correct response is given a score of 1 and an incorrect response is given a score of 0. A participant's performance on the task is calculated by summing scores for each item. Scores range between 0-20, where a higher score indicates greater theory of mind ability. The Hinting Task has been used in a multitude of studies, with the original paper being cited over 1,900 times (Google Scholar).

The Hinting Task has shown good construct validity as a measure of theory of mind in both clinical and non-clinical populations. It has demonstrated clinical utility in its ability to differentiate various clinical populations known to have theory of mind deficits from those without psychopathology, such as autism spectrum disorder (Morrison et al., 2019; Dagdelen, 2020; Saban-Bezalel 2019), bipolar disorder (Bora et al., 2005; Samamé et al., 2015), pediatric bipolar disorder (Schenkel et al., 2008), attention-deficit/hyperactivity disorder (Dagdelen, 2020), schizophrenia (Bora et al., 2009; Pinkham et al., 2016; Braak et al., 2022; Akiyama et al., 2024), first-episode psychosis (Mallawaarachchi et al., 2019), and antisocial personality disorder

(Tasios et al., 2024). Higher scores on the Hinting Task are also significantly associated with better performance with other aspects of theory of mind such as basic facial emotion recognition (Pictures of Facial Affect Task; $r = .44$; Frøyhaug et al., 2019), complex mental state attribution (Reading the Mind in the Eyes Task; $r = .47$; Tasios et al., 2024), recognition of faux pas (Faux Pas Recognition Test; $r = .58$), and self-reported ability to recognize and identify with the thoughts and feelings of others (Empathy Quotient; $r = .48$; Tasios et al., 2024). The Hinting Task also significantly correlates with functional outcomes like functional capacity (UCSD Performance-Based Skills Assessment; $r = .40$) and interpersonal functioning (Social Skills Performance Assessment; $r = .44$; Pinkham et al., 2018), as well as schizophrenia symptom severity as measured by the Positive and Negative Syndrome Scale ($r = .50$; Tasios et al., 2024). Taken together, the Hinting Task has demonstrated good convergent validity, suggesting that task performance adequately captures individual differences in theory of mind.

Despite demonstrating construct validity, the reliability of Hinting Task scores has been questioned. Evidence of acceptable reliability is summarized in a recent meta-analysis, which found that the task's internal consistency is satisfactory in populations with schizophrenia and autism spectrum disorder ($\alpha = .71$ and .77, respectively) but poor in non-clinical populations ($\alpha = .55$; Tsui et al., 2024a). Meta-analytic evidence also suggests that the task has poor test-retest reliability in both samples with schizophrenia of ($r = .65$) and non-clinical samples ($r = .53$; Tsui et al., 2024a). Another factor limiting confidence in the reliability of Hinting Task scores is the presence of ceiling effects. At initial testing, 8% of patients with schizophrenia, 20.3% of matched controls without a psychiatric diagnosis, and 21.9% of an undergraduate sample scored at ceiling (Klein et al., 2020, 2024). Taken together, the Hinting Task generally displays inadequate reliability and problematic ceiling effects, particularly in non-clinical samples.

The suboptimal psychometrics of Hinting Task performance has led researchers to implement a more stringent and standardized scoring criteria (See Figure 1; Klein et al., 2020; Pinkham et al., 2018). Whereas the original scoring criteria allowed for a wide range of responses to receive full points, these revised criteria required certain words or subjects to be mentioned to merit a perfect score on an item. A response must not only describe what the speaker wants, but also explicitly state that the speaker would like the other character in the story to fulfill the desire. A second, more minor change to the scoring criteria is that raters are permitted to assign a score of 1 point after the first hint without issuing a second hint if the initial response lies between the criteria of a 0 and 2. These changes were specifically designed to mitigate ceiling effects and minimize scoring ambiguity, thereby enhancing the psychometric properties of the task (Klein et al., 2020).

Employing their revised scoring criteria, the authors reported that internal consistency did not meet conventional thresholds for acceptability ($\alpha = .70$) in either the patient ($\alpha = .68$) or control group ($\alpha = .64$); however, ceiling effects were substantially reduced. The only notable ceiling effect was in the control sample at the second administration of the Hinting Task, where 8% of respondents received a perfect score. The authors attributed this significant reduction in ceiling effects compared to the findings of previous literature to the more stringent nature of the revised scoring criteria. The authors concluded that overall, the Hinting Task with the revised scoring criteria demonstrated acceptable properties and they recommended it for use in clinical trials (Pinkham et al., 2018).

Only a single study to date has directly compared the original and revised Hinting Task scoring criteria (Klein et al., 2020).[1] The study included patients with schizophrenia, patients

[1] The sample for this study largely overlaps with the sample used in the study that introduced the modified scoring criteria (Pinkham et al., 2018).

with early psychosis, and matched control samples without any psychiatric diagnoses. The revised scoring criteria resulted in significantly lower ceiling effects relative to the original scoring criteria across all groups (<11% vs <31%, respectively). Further, the revised criteria differentiated the schizophrenia group from the control group just as well as the original scoring criteria (d = 0.77 vs 0.79) and better differentiated the early psychosis group from the control sample (d = 0.81 vs 0.59). The revised criteria also resulted in significantly stronger correlations between the Hinting Task and expected outcome measures such as functional capacity as compared to the original in the schizophrenia group (rs = .38 vs .28). Nevertheless, not all psychometric properties were improved with the revised scoring criteria. Internal consistency was suboptimal across all groups and did not significantly change across scoring methods (αs < .69). Overall, the authors concluded that the revised scoring method provides unique psychometric benefits with limited drawbacks as compared to the original scoring system, and thus they strongly recommended a wider adoption of these criteria when administering the Hinting Task (Klein et al., 2020).

To date, few published research articles have reported data from the Hinting Task using the revised criteria without significant additional modifications (e.g., translations, reducing items). Of these studies, all have been in non-psychiatric or psychosis-related samples. Hinting Task scores derived using the revised criteria continue to successfully differentiate psychotic or psychosis-prone groups from those without psychopathology and correlate with psychosis proneness (Wastler & Lenzenweger, 2019; Tsui et al., 2024b). Internal consistency estimates were found to be acceptable in patients with psychosis (αs > .78) but suboptimal in healthy controls (α = .66; Tsui et al., 2024b).

Despite the use of the revised criteria in the abovementioned literature, two aspects of this scoring method have not yet been evaluated. First, the revised scoring method has not been used or validated in clinical populations outside of those with schizophrenia spectrum disorders. Given that theory of mind deficits are implicated across nearly all forms of mental illness (see Cotter et al., 2018), research is needed to determine whether the revised criteria improve the psychometric properties of Hinting Task performance relative to the original scoring criteria in other clinical populations. In the current study, we examine the ceiling effects and convergent validity of Hinting Task performance. Consistent with prior literature, we focused on convergent validity with psychotic symptoms, performance on another measure of theory of mind, and interpersonal functioning. In addition, we also assessed convergent validity with depressive symptoms, given that depressive symptoms are implicated across nearly all forms of psychopathology (Guineau et al., 2023; Soda et al., 2024) and have been robustly linked to theory of mind impairments (Bora et al., 2016; Nestor et al., 2022). Knowledge regarding the transdiagnostic comparative reliability and validity of the Hinting Task scoring criteria will allow researchers to make informed decisions on which scoring method will have the strongest psychometric properties when using the task in clinically heterogenous samples.

Second, research to date has not examined *why* the revised Hinting Task scoring criteria results in worse performance relative to the original scoring criteria. The assumption underlying the use of the revised criteria is that the resulting scores demonstrate greater construct validity with theory of mind compared to the original (Klein et al., 2020). However, an alternative explanation is that the revised scoring criteria introduce verbosity as a confound variable. The revised criteria punish responses lacking specific words or subjects which are generally not necessary to indicate theory of mind. While a less verbose participant might fully understand the

speaker's intention, they may not explicitly acknowledge all aspects of it in their response (e.g., that the intention belongs to the speaker and it is directed at the other character), causing the revised criteria to assign them a lower score compared to the original criteria. As such, individual differences in scores may not be due not to a difference in accuracy of theory of mind per say, but rather due to the amount of speech produced by the participant in their response.

Diminished speech output, or "alogia," is one of the negative symptoms present in schizophrenia spectrum and other psychotic disorders (DSM-5). In psychotic samples, scores resulting from the revised criteria may be uniquely related to functional outcomes because these scores are measuring not only theory of mind, but also alogia, a symptom of schizophrenia. Though the introduction of this confound variable may not have affected the perceived validity of the revised criteria in previous studies focused on populations with schizophrenia spectrum disorders, it may reduce the construct validity of Hinting Task scores as a measure of theory of mind. To summarize, the revised criteria, which penalizes brevity, may not reduce the ability of the Hinting Task to discriminate between psychotic participants (who often experience alogia) and non-clinical individuals, but could potentially reduce the validity of the task as a theory of mind measure in clinical populations not characterized by alogia. In the current study, we test the novel hypothesis that participant verbosity may explain differences in scores across scoring criteria.

The overarching goal of the current study was to compare the utility of the original and revised scoring methods for Hinting Task scores as a measure of theory of mind in a heterogenous clinical sample. The first research aim was to compare the psychometric properties of the original Hinting Task scoring criteria to the revised scoring criteria in a diagnostically diverse sample with severe and complex psychopathology. Consistent with Klein and colleague's

findings in psychotic samples (2020), we hypothesized that participant scores would be significantly lower when applying the revised criteria compared to the original criteria. Further, we hypothesized that revised criteria would result in lower ceiling effects relative to the original scoring method. Finally, we hypothesized that, across both scoring methods, worse performance on the Hinting Task would be significantly associated with worse performance on another theory of mind task, more psychotic symptoms, more severe depressive symptoms, and worse interpersonal functioning. However, we hypothesized that Hinting Task performance would interact with scoring method to predict psychotic symptoms, such that lower Hinting Task performance would have a stronger association with more severe psychotic symptoms using the revised scoring criteria compared to the original criteria.

Our second aim was to determine whether verbosity helps to explain the differences in Hinting Task scores between scoring criteria. First, we hypothesized that higher word count would be associated with better Hinting Task performance while controlling for scoring method. We also hypothesized that the effect of scoring method on Hinting Task performance would depend on word count, such that the differences in performance across scoring criteria would be greater for participants with lower word count relative to those with higher word count.

Method

Transparency and Openness

We report how we determined our sample size, all data exclusions, and all measures in the study, and we follow Journal Article Reporting Standards (Kazak, 2018). This study's hypotheses and analyses were pre-registered. This preregistration as well as all deidentified data, syntax, and results from this study are available on the OSF website for this project: https://osf.io/24jgv/.

Participants

Participants were patients seeking psychiatric treatment at McLean Hospital's Behavioral Health Partial (BHP) Program in the Northeastern United States. Inclusion criteria were as follows: (1) 18 years or older (2) English-speaking (3) assigned to a therapist taking part in the study. This data is being collected as part of a larger study on the effects of cognitive-behavioral therapy on theory of mind (Hudson et al., 2024). Our sample consisted of 173 participants, ranging in age from 19 to 74 years (M = 34.4, SD = 12.9). There were 113 females and 60 males. The majority of participants identified as Non-Hispanic White (80%), followed by Asian (7%), Hispanic White (5%), Multiracial (4%), and Black (3%). One participant indicated that they did not know their race. Participants were generally highly educated, with 38% having received post-college education, 28% holding a 4-year college degree, and 28% having completed some college. A structured clinical interview was not administered, though patient medical records were used to characterize the sample (See Table 1).[2] A priori power analyses suggested that a sample size of 160 participants was required to detect our estimated effect sizes with 80% power.[3] As such, we were well-powered to test our hypotheses.

Measures

Theory of Mind

Hinting Task. Participants are read aloud 10 short vignettes describing an interaction between two characters. Each passage ends with one of the characters dropping a "hint." Participants are asked to state what that character really means. Correct responses are awarded 2 points, with a maximum total score of 20 possible points. Incorrect responses prompt the experimenter to read the second hint, after which participants are asked to state what one

[2] Three participants had received psychotic-spectrum diagnoses. Removing these data points did not change the pattern of results for all analyses.
[3] See Table 2 for details on power analysis results

character wants the other character to do. A correct response after the second hint provides 1 point and an incorrect response provides no points. The Hinting Task has demonstrated convergent validity as a theory of mind task, though internal consistency estimates have previously been reported below acceptable threshold (Klein et al., 2020; Pinkham et al., 2018). In the current study, internal consistency (α) was .52 for the original scoring criteria and .56 for the revised scoring criteria.

Reading the Mind in the Eyes Test. To assess convergent validity, the Reading the Mind in the Eyes Task (RMET; Baron-Cohen et al., 2001) was used to measure the emotion recognition component theory of mind. Participants viewed 36 photographs of faces with only the eye regions visible. Participants were instructed to select which of four complex mental states best matched each photo. Scores were calculated as the percentage of items answered correctly, with higher scores indicating greater theory of mind ability. The RMET has shown acceptable internal consistency and test-retest reliability and demonstrates convergent validity with other measures of theory of mind (Fossati et al., 2017; Murphy & Hall, 2024). In the current sample, internal consistency (α) was .59.

Psychiatric Symptom Severity

Psychotic symptoms. Severity of psychotic symptoms was measured by the Prodromal Questionnaire-Brief (PQ-B; Loewy et al., 2011). This 21-item questionnaire assesses positive symptoms associated with schizophrenia disorders and is used to screen for prodromal psychosis. It has been shown to demonstrate strong internal consistency and concurrent validity with interview-based measures of prodromal syndromes (Loewy et al., 2011). In the current sample, internal consistency was good (α = .87).

Depressive symptoms. Depressive symptom severity was measured with the Patient Health Questionnaire (PHQ-9; Kroenke et al., 2001). The PHQ-9 is a 9-item self-report questionnaire that assesses multiple aspects of depression symptomatology including anhedonia, sleep disturbance, change in appetite, and suicidality. It has demonstrated good internal consistency and strong convergent validity when used in a partial hospital setting (Beard et al., 2016). In the current sample, internal consistency was good ($\alpha = .84$).

Interpersonal Functioning

Interpersonal functioning was assessed by the Behavior and Symptom Identification Scale—Relationships subscale (BASIS-24; Eisen et al., 2004). This subscale is composed of 5 items that assess the quality of relationships, success in social situations, feelings of closeness to others, confidence, and support as reported by the participant in the past week. Items were rated on a 5-point Likert scale, with higher numbers indicating poorer interpersonal functioning. The BASIS-24 has demonstrated good internal consistency, sensitivity, and concurrent validity with other measures of functioning across multiple racial/ethnic groups (Eisen et al., 2004, 2006). In the current sample, internal consistency was acceptable ($\alpha = .76$).

Procedures

At admission to the program (day 1 or 2), participants completed measures of theory of mind, psychotic symptom severity, depression symptom severity, and interpersonal functioning. These measures were completed on participants' own devices, with the exception of the RMET which was administered in-person on a hospital computer using the software E-Prime (Psychology Software Tools, 2023) and the Hinting Task, which was conducted verbally with an experimenter.

Hinting Task data was scored at the time of administration using the revised scoring criteria. Task administrations were audio recorded. To obtain scores using the original scoring methods, each interview was reviewed by two independent raters who were blind to the results of the initial scoring. The raters were also blind to the criteria of the revised scoring method due to the bias that this could introduce to their ratings using the original criteria.

Verbosity was assessed by counting the number of words a participant uttered in their responses during the administration of the Hinting Task. We first transcribed recordings of participants' speech during the Hinting Task using OpenAI's Whisper software (OpenAI, 2022). We then removed all utterances made by the experimenter. Because the number of prompts that participants received was variable depending on the accuracy of their initial responses, we also removed participant responses to secondary hints such that only responses to the first hint were counted towards the verbosity measure. In addition, only task-relevant responses were counted towards the verbosity measure—clarifying questions and non-task-related comments were excluded.

Statistical Analysis

All statistical analyses were performed using R Statistical Software (v4.5.1; R Core Team 2025) using the *tidyverse* (Wickham et al., 2019)*, irr* (Gamer & Lemon, 2019), *bestNormalize* (Peterson, 2021), *ggeffects* (Lüdecke 2018)*, lme4* (Bates et al., 2015)*, lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2017), *effectsize* (Ben-Shachar et al., 2020), *parameters* (Lüdecke et al., 2020), and *interactions* packages **(**Long, 2024). To compare mean score differences across the original and revised scoring criteria, we conducted a paired t-test. To compare ceiling effects, we performed a McNemar's test to determine if one set of criteria had significantly more participants obtaining perfect scores than the other. To assess the convergent validity of Hinting

Task scores, we conducted a series of multiple linear regression models with Hinting Task performance as the predictor and RMET scores, psychotic symptoms, depression severity, and interpersonal functioning as outcome variables. These regressions were conducted separately for each scoring criteria. Age, sex, and ethnicity were included as covariates. We accounted for multiple comparisons by employing the false discovery rate procedure (Benjamini & Hochberg, 1995).

To investigate whether the Hinting Task's association with psychotic symptom severity differed in strength depending on scoring criteria, we conducted a linear regression with Hinting Task performance, scoring criteria, and their interaction term as predictors and psychotic symptom severity as the outcome variable. In addition, we deviated from our preregistration to conduct additional post hoc regression analyses examining the interaction between Hinting Task performance and scoring criteria predicting RMET performance, depression severity, and interpersonal functioning to determine whether the convergent validity of Hinting Task performance varied as a function of scoring criteria. Age, sex, and ethnicity were again included as covariates and multiple comparisons were accounted for by employing the false discovery rate procedure.

To assess the role of verbosity in Hinting Task performance, we first performed a multilevel regression with scoring method and word count as the predictor variables and Hinting Task performance as the outcome variable. Next, we added the interaction between scoring criteria and word count in the model to determine whether the impact of scoring criteria on performance varies as a function of participant verbosity. Multilevel models were selected to

account for between-person differences due to the nested nature of Hinting Task scores within participants.[4] Age, sex, and ethnicity were again included as covariates.

Results

Preliminary Analyses

Bivariate correlations are presented in Table 3. The means, standard deviations, and associations of primary study variables with demographic variables are depicted in Table 4. Demographic variables that were included as covariates in analyses showed no significant associations with any outcome variables, with the exception of females reporting higher depressive symptoms on the PHQ-9 than males.

To assess inter-rater reliability, the intraclass correlation coefficient (ICC) was calculated for Hinting Task scores using both scoring criteria. The ICC was good (ICC [1,3] = .82) using the original criteria and excellent (ICC [1,2] = .93) using the revised criteria.

Aim 1: Comparison of Psychometric Properties Across Scoring Criteria

We first investigated the mean Hinting Task scores resulting from both the original and revised scoring criteria. Consistent with our hypothesis, Hinting Task scores were significantly lower when applying the revised scoring criteria as compared to the original scoring criteria, reflecting a large effect size (d = 0.80, 95% CI [0.62, 0.98]; Original: M = 16.85, SD = 2.43; Revised: M = 15.36, SD = 2.23; $p < .001$). See Figure 2 for a visual comparison of these mean values. Also consistent with our hypotheses, ceiling effects (i.e., the proportion of participants with a perfect score) were significantly lower for Hinting Task scores obtained using the revised criteria as compared to the original criteria (1.22% vs 12.80%, respectively; $\chi^2(1) = 15.43$, $p < .001$).

[4] Although our preregistration stated that we would model random intercepts *and* random slopes, we deviated from this plan and only modeled random intercepts due to model convergence issues. Random slopes cannot be meaningfully estimated with only two within-person observations.

We assessed whether scoring criteria affected the convergent validity of Hinting Task performance with inference of complex mental states from images of eyes (i.e., another measure of theory of mind; RMET), psychiatric symptom severity (i.e., psychotic symptoms [PQ-B] and depressive symptoms [PHQ-9]), and interpersonal functioning (i.e., BASIS-24 Relationships subscale). The interaction between Hinting Task performance and scoring criteria was not significant across each of these analyses (see Table 5 for effect sizes), suggesting that the convergent validity of Hinting Task performance did not vary as a function of scoring criteria. This finding is inconsistent with our hypothesis that Hinting Task performance derived from the revised criteria would be a significantly stronger predictor of psychotic symptoms than scores derived using the original criteria.

Consistent with our hypotheses, Hinting Task performance demonstrated convergent validity with the RMET for both the original and revised scoring criteria (see Table 5). Partially consistent with our hypotheses, Hinting Task performance using the original criteria (but not revised criteria) was significantly associated with psychotic symptoms (β = -.18, 95% CI [-.33, -.02] p = .03; β = -.12, 95% CI [-.28, .03], p = .12, respectively); however, these effects were not statistically different, and the effect was no longer significant when correcting for multiple comparisons (see Table 5). Inconsistent with hypotheses, Hinting Task performance was not significantly associated with depression severity or interpersonal functioning using either scoring criteria.

Aim 2: Investigation of the Role of Verbosity in Hinting Task Performance

We first assessed the main effects of scoring criteria and verbosity on Hinting Task performance. Before outliers were removed, we found that when controlling for scoring criteria, verbosity was significantly associated with Hinting Task performance (β = .14, 95% CI [.00,

.27], p = .045); however, after outliers were removed, this effect was no longer significant (β = .10, 95% CI [-.04, .23], p = .17). When controlling for verbosity, scoring criteria was significantly associated with Hinting Task performance (β = -.55, 95% CI [-.74, -.52], p < .001). These main effects were qualified by a significant scoring criterion by verbosity interaction. Consistent with hypotheses, the effect of scoring criteria on Hinting Task performance was dependent on verbosity (β = .16, 95% CI [.05, .27], p = .002). Follow-up analyses indicated the magnitude of the effect of scoring criteria on performance was greater for participants with below average verbosity (β = -.79, 95% CI [-.94, -.63], p < .001) relative to those with average (β = -.63, 95% CI [-.74, -.52], p < .001) or above average verbosity (β = -.48, 95% CI [-.63, -.32], p < .001). See Figure 3 for a visualization of the interaction between scoring criteria and verbosity predicting Hinting Task performance.

Discussion

The present study was the first to evaluate the psychometric properties of the Hinting Task using the original and revised scoring criteria in a clinically heterogeneous sample. Consistent with hypotheses, results indicated that the revised scoring criteria resulted in significantly lower scores and reduced ceiling effects as compared to the original scoring criteria. Partially consistent with hypotheses, we found that Hinting Task performance, regardless of scoring criteria, was associated with performance on another measure of theory of mind, but not psychiatric symptom severity or interpersonal functioning. We also tested the novel hypothesis that participant verbosity may help to explain differences in task scores between scoring criteria. Our results supported this hypothesis, indicating that participants who were less verbose tended to have larger discrepancies in their performance between the different scoring criteria.

Consistent with hypotheses and prior research (Klein et al., 2020), we found that the revised scoring criteria resulted in significantly lower scores and reduced ceiling effects as compared to the original scoring criteria. These results suggest that the revised criteria provides a more stringent bar for what is considered a correct response to items on the Hinting Task. Whereas the original criteria were loosely defined and provided room for raters to give participants the benefit of the doubt on a questionable response, the revised criteria provide strict requirements that must be met to receive full points on a given item. This study is the first to show that these differences between the scoring criteria extend to clinical populations outside of psychotic and non-clinical samples. Researchers can expect these lower scores and reduced ceiling effects from the revised criteria to generalize to a wide range of clinical populations.

Large ceiling effects artificially restrict the variance of a measure, limiting the strength of its association with theoretically related constructs. Klein et al., 2020 concluded that the revised scoring criteria provided a more valid estimate of theory of mind ability than the original criteria in part due to the reduced ceiling effects. While we also observed this difference in ceiling effects in a diagnostically diverse clinical sample, we found that the revised criteria did not improve the convergent validity of the Hinting Task compared to the original criteria. Despite removal of a problematic range-restricting effect, the revised criteria do not covary more with relevant outcome variables than the original criteria.

One potential explanation for why the revised scoring criteria reduces ceiling effects but does not improve validity is that it introduces verbosity as a confounding variable. Indeed, our results suggest that verbosity uniquely predicts participant performance on the Hinting Task under the revised criteria. The revised criteria disadvantage participants with lower verbosity, more heavily penalizing those who are more succinct than those who are more verbose. While

this confound reduces ceiling effects, our results suggest that it does not improve the convergent validity of Hinting Task performance. We predict that future researchers may only expect to see improved convergent validity in outcomes that are associated with verbosity. In sum, our results suggest that the use of the revised criteria artificially improves ceiling effects while contaminating the construct validity of the Hinting Task performance.

This study is the first to evaluate the psychometric properties of the Hinting Task using the revised scoring criteria in a clinical sample that was not characterized by psychotic disorders. Klein and colleagues (2020) strongly recommended the use of the revised scoring system after comparing criteria in schizophrenic, early psychosis, and non-clinical samples, citing improved psychometric properties. However, the results of this study indicate a less clear-cut comparison of the scoring criteria when measuring theory of mind in clinical samples not characterized by psychosis. The revised criteria continues to be an effective remedy to the problematic ceiling effects that often accompany Hinting Task scores using the original criteria; however, it does so at the expense of construct validity. More specifically, the revised scoring method introduces construct contamination through its association with participant verbosity. This association with verbosity would explain why Klein and colleagues (2020) found that implementation of the revised criteria better differentiates early psychosis from control participants, given the presence of negative symptoms such as alogia (reduced speech). If researchers wish to use the Hinting Task to help differentiate psychotic-spectrum from non-clinical populations, scoring responses using the revised criteria may provide more utility, as it is measuring two constructs implicated in psychosis (theory of mind ability and verbosity). However, if researchers wish to use the Hinting Task as a social cognition or theory of mind measure, implementation of the revised

criteria should be done cautiously. Verbosity should be included as a covariate in analyses when the other variables of interest may be associated with the construct.

Inconsistent with past research, we did not find a significant association between Hinting Task performance and psychiatric (i.e., psychotic and depressive) symptoms or interpersonal functioning. These findings may be due to the characteristics of our sample and measures. For example, our sample was not characterized by psychotic disorders. As such, the vast majority of any psychotic symptoms that were reported may have been a product of other presenting concerns (e.g., social anxiety can lead to endorsement of low-level paranoia that is more characteristic of anxiety than psychosis; Freeman et al., 2008). In addition, although evidence suggests that theory of mind ability is associated with depressive symptoms (Bora et al., 2016; Nestor et al., 2022), these findings often result from samples with a wide range of depressive symptoms. Given that our sample focused on a severe clinical population in which most participants endorsed significant depressive symptoms, our restricted range may have precluded us from detecting these effects. Finally, we assessed interpersonal functioning with a self-report questionnaire that asked about relationship satisfaction, rather than a performance-based assessment that rates competency of social skills often used in prior research (Klein et al., 2020; Pinkham et al., 2016); this discrepancy may explain its lack of association with the Hinting Task in the current study.

This study had a number of important strengths that bolster confidence in our conclusions. First, the study was pre-registered and adequately powered to detect the hypothesized effects. Additionally, the other theory of mind measure we implemented (RMET) is a task-based measure, limiting potential bias that often arises from self-reported theory of mind assessment (Murphy & Lilienfeld, 2019). Blinding procedures were enacted such that raters of

each scoring criteria were independent and not privy to the scores obtained using the alternative criteria or any other participant data. Finally, participants were recruited from a treatment setting, ensuring that the sample was clinically representative with a wide range of presenting concerns and comorbidities, increasing the generalizability of our findings.

In addition to these strengths, the results must be considered in light of the following limitations. While the theory of mind measures were performance-based, many of the other outcome measures (i.e., symptom severity, interpersonal functioning) were obtained through self-report questionnaires that introduce potential bias due to demand characteristics or limited insight. Additionally, the sample was relatively racially homogeneous, limiting the ability to generalize results to non-White individuals. Finally, both of our theory of mind measures demonstrated poor internal consistency. This finding suggests that these measures may not provide sufficiently reliable estimates of individual differences in theory of mind ability. Low internal consistency is common for performance-based tasks with relatively few items (Revelle, 2024) and may improve if these tasks included additional items. Regardless, the limited reliability of Hinting Task scores across both scoring approaches may have attenuated their associations with related constructs, underscoring the need for more robust and well-validated measures of theory of mind in future research.

The current study demonstrated that within a heterogeneous clinical sample, the revised scoring criteria for the Hinting Task resulted in lower scores and reduced ceiling effects relative to the original criteria. These benefits come at a cost: The revised criteria also introduced participant verbosity as a confound without improving convergent validity over the original criteria. Although past research suggests that the implementation of the revised criteria may provide additional utility in predicting functional capacity and differentiating psychotic-spectrum

populations from those without psychotic disorders (Klein et al., 2020), our research suggests that it largely does not demonstrate superior psychometric properties in a non-psychotic clinical sample and contaminates construct validity through its association with participant verbosity. As such, we recommend that researchers who implement the revised criteria include verbosity as a covariate to preserve construct validity when it is associated with the other variables of interest. Importantly, concerns about the psychometric validity of theory of mind measurement extend beyond the Hinting Task (Quesque & Rossetti, 2020; Yeung et al., 2024), and the field is in need of more valid and reliable theory of mind assessment tools. Given the purported centrality of theory of mind as a mechanism underlying psychopathology and interpersonal functioning (Cotter et al., 2018; Gur & Gur, 2016), accurate assessment is crucial for advancing theoretical models and informing clinical care.

References

Akiyama, H., Okubo, R., Toyomaki, A., Miyazaki, A., Hattori, S., Nohara, M., Sasaki, Y., Kubota, R., Okano, H., Takahashi, K., Hasegawa, Y., Wada, I., Uchino, T., Takeda, K., Ikezawa, S., Nemoto, T., Ito, Y. M., & Hashimoto, N. (2024). The evaluation study for social cognition measures in Japan: Psychometric properties, relationships with social function, and recommendations. *Asian Journal of Psychiatry*, *95*, 104003. https://doi.org/10.1016/j.ajp.2024.104003

Badcock, P. B., Davey, C. G., Whittle, S., Allen, N. B., & Friston, K. J. (2017). The Depressed Brain: An Evolutionary Systems Theory. *Trends in Cognitive Sciences*, *21*(3), 182–194. https://doi.org/10.1016/j.tics.2017.01.005

Bateman, A., & Fonagy, P. (2013). Mentalization-Based Treatment. *Psychoanalytic Inquiry*, *33*(6), 595–613. https://doi.org/10.1080/07351690.2013.835170

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Beard, C., Hsu, K. J., Rifkin, L. S., Busch, A. B., & Björgvinsson, T. (2016). Validation of the PHQ-9 in a psychiatric sample. *Journal of Affective Disorders*, *193*, 267–273. https://doi.org/10.1016/j.jad.2015.12.075

Ben-Shachar M, Lüdecke D, Makowski D (2020). effectsize: Estimation of Effect Size Indices and Standardized Parameters. Journal of Open Source Software, 5(56), 2815. doi: 10.21105/joss.02815

Blevins, C. A., Weathers, F. W., Davis, M. T., Witte, T. K., & Domino, J. L. (2015). The Posttraumatic Stress Disorder Checklist for *DSM□5* (PCL□5): Development and Initial Psychometric Evaluation. *Journal of Traumatic Stress*, *28*(6), 489–498. https://doi.org/10.1002/jts.22059

Bora, E., Bartholomeusz, C., & Pantelis, C. (2016). Meta-analysis of Theory of Mind (ToM) impairment in bipolar disorder. *Psychological Medicine*, *46*(2), 253–264. https://doi.org/10.1017/S0033291715001993

Bora, E., Vahip, S., Gonul, A. S., Akdeniz, F., Alkan, M., Ogut, M., & Eryavuz, A. (2005). Evidence for theory of mind deficits in euthymic patients with bipolar disorder. *Acta Psychiatrica Scandinavica*, *112*(2), 110–116. https://doi.org/10.1111/j.1600-0447.2005.00570.x

Bora, E., Yucel, M., & Pantelis, C. (2009). Theory of mind impairment in schizophrenia: meta-analysis. *Schizophrenia research*, *109*(1-3), 1-9.

Braak, S., Su, T., Krudop, W., Pijnenburg, Y. A. L., Reus, L. M., Van Der Wee, N., Bilderbeck, A. C., Dawson, G. R., Van Rossum, I. W.-, Campos, A. V., Arango, C., Saris, I. M. J., Kas, M. J., & Penninx, B. W. J. H. (2022). Theory of Mind and social functioning among neuropsychiatric disorders: A transdiagnostic study. *European Neuropsychopharmacology*, *64*, 19–29. https://doi.org/10.1016/j.euroneuro.2022.08.005

Cacioppo, S., Grippo, A. J., London, S., Goossens, L., & Cacioppo, J. T. (2015). Loneliness: Clinical Import and Interventions. *Perspectives on Psychological Science*, *10*(2), 238–249. https://doi.org/10.1177/1745691615570616

Corcoran, R., Mercer, G., & Frith, C. D. (1995). Schizophrenia, symptomatology and social inference: Investigating "theory of mind" in people with schizophrenia. *Schizophrenia Research*, *17*(1), 5–13. https://doi.org/10.1016/0920-9964(95)00024-G

Cotter, J., Granger, K., Backx, R., Hobbs, M., Looi, C. Y., & Barnett, J. H. (2018). Social cognitive dysfunction as a clinical marker: A systematic review of meta-analyses across 30 clinical conditions. *Neuroscience & Biobehavioral Reviews*, *84*, 92–99. https://doi.org/10.1016/j.neubiorev.2017.11.014

Couture, S. M. (2006). The Functional Significance of Social Cognition in Schizophrenia: A Review. *Schizophrenia Bulletin*, *32*(Supplement 1), S44–S63. https://doi.org/10.1093/schbul/sbl029

Cullen, C., Gaynor, K., & Kessler, K. (2025). Evaluation of a brief online multi-index assessment for predicting increased psychotic-like experiences in the community: A perceptual, cognitive and affective approach. *Schizophrenia Research: Cognition*, *40*, 100357. https://doi.org/10.1016/j.scog.2025.100357

Dagdelen, F. (2020). Comparison of social cognition in adolescents diagnosed with attention deficit hyperactivity disorder and autism spectrum disorder. *Dusunen Adam:The Journal of Psychiatry and Neurological Sciences*. https://doi.org/10.14744/DAJPNS.2020.00093

Diedenhofen, B. & Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. PLoS ONE, 10(4): e0121945. doi: 10.1371/journal.pone.0121945 Available: http://dx.doi.org/10.1371/journal.pone.0121945

Dodell-Feder, D., Tully, L. M., Lincoln, S. H., & Hooker, C. I. (2014). The neural basis of theory of mind and its relationship to social functioning and social anhedonia in individuals with

schizophrenia. *NeuroImage: Clinical*, *4*, 154–163. https://doi.org/10.1016/j.nicl.2013.11.006

Dvash, J., & Shamay-Tsoory, S. G. (2014). Theory of Mind and Empathy as Multidimensional Constructs: Neurological Foundations. *Topics in Language Disorders*, *34*(4), 282–295. https://doi.org/10.1097/TLD.0000000000000040

Eisen, S. V., Gerena, M., Ranganathan, G., Esch, D., & Idiculla, T. (2006). Reliability and Validity of the BASIS-24© Mental Health Survey for Whites, African-Americans, and Latinos. *The Journal of Behavioral Health Services & Research*, *33*(3), 304–323. https://doi.org/10.1007/s11414-006-9025-3

Eisen, S. V., Normand, S.-L., Belanger, A. J., Spiro, A., & Esch, D. (2004). The Revised Behavior and Symptom Identification Scale (BASIS-R): Reliability and Validity. *Medical Care*, *42*(12), 1230–1241. https://doi.org/10.1097/00005650-200412000-00010

Fletcher TD (2022). _QuantPsyc: Quantitative Psychology Tools_. R package version 1.6, https://CRAN.R-project.org/package=QuantPsyc.

Fossati, A., Borroni, S., Dziobek, I., Fonagy, P., & Somma, A. (2018). Thinking about assessment: Further evidence of the validity of the Movie for the Assessment of Social Cognition as a measure of mentalistic abilities. *Psychoanalytic Psychology*, 35(1), 127.

Franken, I. H. A., Rassin, E., & Muris, P. (2007). The assessment of anhedonia in clinical and non-clinical populations: Further validation of the Snaith–Hamilton Pleasure Scale (SHAPS). *Journal of Affective Disorders*, *99*(1–3), 83–89. https://doi.org/10.1016/j.jad.2006.08.020

Frith, C. D., & Frith, U. (2008). Implicit and explicit processes in social cognition. *Neuron*, *60*(3), 503-510.

Frøyhaug, M., Andersson, S., Andreassen, O. A., Ueland, T., & Vaskinn, A. (2019). Theory of mind in schizophrenia and bipolar disorder: psychometric properties of the Norwegian version of the Hinting Task. *Cognitive Neuropsychiatry*, *24*(6), 454-469.

Gamer M, Lemon J, <puspendra.pusp22@gmail.com> IFPS (2019). _irr: Various Coefficients of Interrater Reliability and Agreement_. doi:10.32614/CRAN.package.irr https://doi.org/10.32614/CRAN.package.irr, R package version 0.84.1, <https://CRAN.R-project.org/package=irr>.

Green, M. F., & Leitman, D. I. (2008). Social cognition in schizophrenia. *Schizophrenia bulletin*, *34*(4), 670-672.

Greig, T. C., Bryson, G. J., & Bell, M. D. (2004). Theory of Mind Performance in Schizophrenia: Diagnostic, Symptom, and Neuropsychological Correlates. *Journal of Nervous & Mental Disease*, *192*(1), 12–18. https://doi.org/10.1097/01.nmd.0000105995.67947.fc

Guineau, M. G., Ikani, N., Rinck, M., Collard, R. M., Van Eijndhoven, P., Tendolkar, I., Schene, A. H., Becker, E. S., & Vrijsen, J. N. (2023). Anhedonia as a transdiagnostic symptom across psychological disorders: A network approach. *Psychological Medicine*, *53*(9), 3908–3919. https://doi.org/10.1017/S0033291722000575

Gur, R. C., & Gur, R. E. (2016). Social cognition as an RDoC domain. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *171*(1), 132–141. https://doi.org/10.1002/ajmg.b.32394

Gutiérrez-Rojas, L., Porras-Segovia, A., Dunne, H., Andrade-González, N., & Cervilla, J. A. (2020). Prevalence and correlates of major depressive disorder: A systematic review. *Brazilian Journal of Psychiatry*, *42*(6), 657–672. https://doi.org/10.1590/1516-4446-2020-0650

Halverson, T. F., Pinkham, A. E., Harvey, P. D., & Penn, D. L. (2022). Brief battery of the Social Cognition Psychometric Evaluation study (BB-SCOPE): Development and validation in schizophrenia spectrum disorders. *Journal of psychiatric research*, *150*, 307-316.

Hudson, C. C., Fan, K., Bockhorst, J., Beard, C. (2024). Changes in Social Cognition During a Cognitive-Behavioral Therapy-Based Partial Hospitalization Program [Preregistration]. OSF. https://osf.io/6jbpf/

Introducing Whisper. (2022, September 21). OpenAI. https://openai.com/index/whisper/

Johannesen, J. K., Fiszdon, J. M., Weinstein, A., Ciosek, D., & Bell, M. D. (2018). The Social Attribution Task - Multiple Choice (SAT-MC): Psychometric comparison with social cognitive measures for schizophrenia research. *Psychiatry Research*, *262*, 154–161. https://doi.org/10.1016/j.psychres.2018.02.011

Kittel, A. F. D., Olderbak, S., & Wilhelm, O. (2022). Sty in the Mind's Eye: A Meta-Analytic Investigation of the Nomological Network and Internal Consistency of the "Reading the Mind in the Eyes" Test. *Assessment*, *29*(5), 872–895. https://doi.org/10.1177/1073191121996469

Klein, H. S., Springfield, C. R., Bass, E., Ludwig, K., Penn, D. L., Harvey, P. D., & Pinkham, A. E. (2020). Measuring mentalizing: A comparison of scoring methods for the hinting task. *International Journal of Methods in Psychiatric Research*, *29*(2), e1827. https://doi.org/10.1002/mpr.1827

Klein, H., Springfield, C. R., & Pinkham, A. E. (2024). Measuring social cognition within the university: The Social Cognition Psychometric Evaluation (SCOPE) battery in an undergraduate sample. *Applied Neuropsychology: Adult*, *31*(5), 866–873. https://doi.org/10.1080/23279095.2022.2082875

Kosutzka, Z., Kralova, M., Kusnirova, A., Papayova, M., Valkovic, P., Csefalvay, Z., & Hajduk, M. (2019). Neurocognitive Predictors of Understanding of Intentions in Parkinson Disease. *Journal of Geriatric Psychiatry and Neurology*, *32*(4), 178–185. https://doi.org/10.1177/0891988719841727

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kruse, E. A., Saxena, A., Shovestul, B. J., Dudek, E. M., Reda, S., Dong, J., Venkataraman, A., Lamberti, J. S., & Dodell-Feder, D. (2024). Training individuals with schizophrenia to gain volitional control of the theory of mind network with real-time fMRI: A pilot study. *Schizophrenia Research: Cognition*, *38*, 100329. https://doi.org/10.1016/j.scog.2024.100329

Kuznetsova A, Brockhoff PB, Christensen RHB (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." Journal of Statistical Software_, *82*(13), 1-26. doi:10.18637/jss.v082.i13 <https://doi.org/10.18637/jss.v082.i13>.

Lindgren, M., Torniainen-Holm, M., Heiskanen, I., Voutilainen, G., Pulkkinen, U., Mehtälä, T., Jokela, M., Kieseppä, T., Suvisaari, J., & Therman, S. (2018). Theory of mind in a first-episode psychosis population using the Hinting Task. *Psychiatry Research*, *263*, 185–192. https://doi.org/10.1016/j.psychres.2018.03.014

Loewy, R. L., Pearson, R., Vinogradov, S., Bearden, C. E., & Cannon, T. D. (2011). Psychosis risk screening with the Prodromal Questionnaire—Brief Version (PQ-B). *Schizophrenia Research*, *129*(1), 42–46. https://doi.org/10.1016/j.schres.2011.03.029

Long JA (2024). interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions. doi:10.32614/CRAN.package.interactions https://doi.org/10.32614/CRAN.package.interactions, R package version 1.2.0, https://cran.r-project.org/package=interactions.

Lüdecke D (2018). "ggeffects: Tidy Data Frames of Marginal Effects from Regression Models." Journal of Open Source Software_, *3*(26), 772. doi:10.21105/joss.00772 <https://doi.org/10.21105/joss.00772>.

Lüdecke D, Ben-Shachar M, Patil I, Makowski D (2020). "Extracting, Computing and Exploring the Parameters of Statistical Models using R." Journal of Open Source Software, 5(53), 2445. doi:10.21105/joss.02445 <https://doi.org/10.21105/joss.02445>.

Mallawaarachchi, S. R., Cotton, S. M., Anderson, J., Killackey, E., & Allott, K. A. (2019). Exploring the use of the Hinting Task in first-episode psychosis. *Cognitive Neuropsychiatry*, *24*(1), 65–79. https://doi.org/10.1080/13546805.2019.1568864

May, K., & Hittner, J. B. (1997). A note on statistics for comparing dependent correlations. *Psychological reports*, *80*(2), 475-480.

Moreno-Amador, B., Piqueras, J. A., Rodríguez-Jiménez, T., Martínez-González, A. E., & Cervin, M. (2023). Measuring symptoms of obsessive-compulsive and related disorders using a single dimensional self-report scale. *Frontiers in Psychiatry*, *14*, 958015. https://doi.org/10.3389/fpsyt.2023.958015

Morrison, K. E., Pinkham, A. E., Kelsven, S., Ludwig, K., Penn, D. L., & Sasson, N. J. (2019). Psychometric Evaluation of Social Cognitive Measures for Adults with Autism. *Autism Research*, *12*(5), 766–778. https://doi.org/10.1002/aur.2084

Murphy, B. A., & Lilienfeld, S. O. (2019). Are self-report cognitive empathy ratings valid proxies for cognitive empathy ability? Negligible meta-analytic relations with behavioral task performance. Psychological assessment, 31(8), 1062.

Murphy, B. A., & Hall, J. A. (2024). How a strong measurement validity review can go astray: A look at and recommendations for future measurement-focused reviews. *Clinical Psychology Review*, *114*, 102506.

Nestor, B. A., Sutherland, S., & Garber, J. (2022). Theory of mind performance in depression: A meta-analysis. *Journal of Affective Disorders*, *303*, 233–244. https://doi.org/10.1016/j.jad.2022.02.028

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: Toward a brief form for research and applied settings. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01503

Peterson, R. A. (2021). Finding Optimal Normalizing Transformations via bestNormalize. The R Journal, 13:1, 310-329, DOI:10.32614/RJ-2021-041

Pinkham, A. E., Harvey, P. D., & Penn, D. L. (2018). Social Cognition Psychometric Evaluation: Results of the Final Validation Study. *Schizophrenia Bulletin*, *44*(4), 737–748. https://doi.org/10.1093/schbul/sbx117

Pinkham, A. E., Penn, D. L., Green, M. F., & Harvey, P. D. (2016). Social Cognition Psychometric Evaluation: Results of the Initial Psychometric Study. *Schizophrenia Bulletin*, *42*(2), 494–504. https://doi.org/10.1093/schbul/sbv056

Psychology Software Tools, Inc. [E-Prime Go]. (2020). Retrieved from https://support.pstnet.com/.

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. Perspectives on Psychological Science, 15(2), 384-396.

Revelle, W. (2024). The seductive beauty of latent variable models: Or why I don't believe in the Easter Bunny. Personality and Individual Differences, 221, 112552.

Rocca, P., Galderisi, S., Rossi, A., Bertolino, A., Rucci, P., Gibertoni, D., Montemagni, C., Bellino, S., Aguglia, E., Amore, M., Bellomo, A., Biondi, M., Carpiniello, B., Cuomo, A., D'Ambrosio, E., dell'Osso, L., Girardi, P., Marchesi, C., Monteleone, P., … Goracci, A. (2018). Disorganization and real-world functioning in schizophrenia: Results from the multicenter study of the Italian Network for Research on Psychoses. *Schizophrenia Research*, *201*, 105–112. https://doi.org/10.1016/j.schres.2018.06.003

Saban-Bezalel, R., Dolfin, D., Laor, N., & Mashal, N. (2019). Irony comprehension and mentalizing ability in children with and without Autism Spectrum Disorder. *Research in Autism Spectrum Disorders*, *58*, 30–38. https://doi.org/10.1016/j.rasd.2018.11.006

Samamé, C., Martino, D. J., & Strejilevich, S. A. (2015). An individual task meta-analysis of social cognition in euthymic bipolar disorders. *Journal of Affective Disorders*, *173*, 146-153.

Saxena, A., Shovestul, B. J., Dudek, E. M., Reda, S., Venkataraman, A., Lamberti, J. S., & Dodell-Feder, D. (2023). Training volitional control of the theory of mind network with real-time fMRI neurofeedback. *NeuroImage*, *279*, 120334. https://doi.org/10.1016/j.neuroimage.2023.120334

Schenkel, L. S., Marlow-O'Connor, M., Moss, M., Sweeney, J. A., & Pavuluri, M. (2008). Theory of mind and social inference in children and adolescents with bipolar disorder. *Psychological medicine*, *38*(6), 791-800.

Snaith, R. P., Hamilton, M., Morley, S., Humayan, A., Hargreaves, D., & Trigwell, P. (1995). A Scale for the Assessment of Hedonic Tone the Snaith–Hamilton Pleasure Scale. *British Journal of Psychiatry*, *167*(1), 99–103. https://doi.org/10.1192/bjp.167.1.99

Soda, T., Toyama, A., Takeda, M., Kunisato, Y., & Yamashita, Y. (2024). *Evaluating the General Psychopathological Factor (p-Factor) using the DSM-5 Level 1 Cross-Cutting Symptom Measure in the General Population*. PsyArXiv. https://doi.org/10.31234/osf.io/6qy84

Tasios, K., Douzenis, A., Gournellis, R., & Michopoulos, I. (2024, January). Empathy and Violence in Schizophrenia and Antisocial Personality Disorder. In *Healthcare* (Vol. 12, No. 1, p. 89). Multidisciplinary Digital Publishing Institute.

Teli, P. K. (n.d.). *Theory of Mind (ToM) in Individuals with Obsessive Compulsive Disorder (OCD), First Degree Relatives and Healthy Controls: An Endophenotype Study.*

Thibaudeau, É., Cellard, C., Turcotte, M., & Achim, A. M. (2021). Functional Impairments and Theory of Mind Deficits in Schizophrenia: A Meta-analysis of the Associations. *Schizophrenia Bulletin*, *47*(3), 695–711. https://doi.org/10.1093/schbul/sbaa182

Tousignant, B., Jackson, P. L., Massicotte, E., Beauchamp, M. H., Achim, A. M., Vera-Estay, E., Bedell, G., & Sirois, K. (2018). Impact of traumatic brain injury on social cognition in adolescents and contribution of other higher order cognitive functions. *Neuropsychological Rehabilitation*, *28*(3), 429–447. https://doi.org/10.1080/09602011.2016.1158114

Trojsi, F., Siciliano, M., Russo, A., Passaniti, C., Femiano, C., Ferrantino, T., De Liguoro, S., Lavorgna, L., Monsurrò, M. R., Tedeschi, G., & Santangelo, G. (2016). Theory of Mind and Its Neuropsychological and Quality of Life Correlates in the Early Stages of Amyotrophic Lateral Sclerosis. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01934

Tsui, H. K. H., Wong, T. Y., Ma, C. F., Wong, T. E., Hsiao, J., & Chan, S. K. W. (2024a). Reliability of Theory of Mind Tasks in Schizophrenia, ASD, and Nonclinical Populations: A Systematic Review and Reliability Generalization Meta-analysis. *Neuropsychology Review*. https://doi.org/10.1007/s11065-024-09652-4

Tsui, H. K. H., Liao, Y., Hsiao, J., Suen, Y. N., Yan, E. W. C., Poon, L. T., ... & Chan, S. K. W. (2024b). Mentalizing impairments and hypermentalizing bias in individuals with first-episode schizophrenia-spectrum disorder and at-risk mental state: the differential roles of neurocognition and social anxiety. *European Archives of Psychiatry and Clinical Neuroscience*, 1-13.

Vellante, M., Baron-Cohen, S., Melis, M., Marrone, M., Petretto, D. R., Masala, C., & Preti, A. (2013). The "Reading the Mind in the Eyes" test: Systematic review of psychometric properties and a validation study in Italy. *Cognitive Neuropsychiatry*, *18*(4), 326–354. https://doi.org/10.1080/13546805.2012.721728

Velthorst, E., Socrates, A., GROUP Investigators, Alizadeh, B. Z., Van Amelsvoort, T., Bartels-Velthuis, A. A., Bruggeman, R., Cahn, W., De Haan, L., Schirmbeck, F., Simons, C. J. P., Van Os, J., & Fett, A.-K. (2023). Age-Related Social Cognitive Performance in

Individuals With Psychotic Disorders and Their First-Degree Relatives. *Schizophrenia Bulletin*, *49*(6), 1460–1469. https://doi.org/10.1093/schbul/sbad069

Wastler, H. M., & Lenzenweger, M. F. (2019). Self-referential hypermentalization in schizotypy. *Personality Disorders: Theory, Research, and Treatment*, *10*(6), 536–544. https://doi.org/10.1037/per0000344

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." _Journal of Open Source Software_, *4*(43), 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.

Yeung, E. K. L., Apperly, I. A., & Devine, R. T. (2024). Measures of individual differences in adult theory of mind: A systematic review. Neuroscience & Biobehavioral Reviews, 157, 105481.

Yilmaz, G., Yildirim, E. A., & Tabakcı, A. S. (2023). Comparison of Social-Evaluative Anxiety and Theory of Mind Functions in Social Anxiety Disorder, Schizophrenia, and Healthy Controls. *Psychopathology*, *56*(6), 440–452. https://doi.org/10.1159/000529880

Zanarini, M. C., Vujanovic, A. A., Parachini, E. A., Boulanger, J. L., Frankenburg, F. R., & Hennen, J. (2003). A Screening Measure for BPD: The McLean Screening Instrument for Borderline Personality Disorder (MSI-BPD). *Journal of Personality Disorders*, *17*(6), 568–573. https://doi.org/10.1521/pedi.17.6.568.25355

Table 1

Clinical Characterization of the Sample from Chart Diagnoses

Diagnostic Category and Specific Diagnoses	***n***	***%***
Neurodevelopmental Disorders	30	17%
Attention-deficit/hyperactivity disorder	30	17%
Autism spectrum disorder	1	1%
Schizophrenia Spectrum and Other Psychotic Disorders	3	2%
Bipolar and Related Disorders	32	18%
Bipolar I disorder	16	9%
Bipolar II disorder	16	9%
Depressive Disorders	130	75%
Major depressive disorder	130	75%
Persistent depressive disorder	3	2%
Anxiety Disorders	60	35%
Generalized anxiety disorder	54	31%
Panic disorder	6	3%
Agoraphobia	4	2%
Social anxiety disorder	3	2%
Other specified anxiety	1	<1%
Obsessive-Compulsive and Related Disorders	9	5%
Trauma- and Stressor-Related Disorders	35	20%
Posttraumatic stress disorder	34	20%
Adjustment disorder	1	<1%
Somatic Symptom and Related Disorders	1	<1%
Feeding and Eating Disorders	12	7%
Anorexia nervosa	3	2%
Bulimia nervosa	2	1%
Binge eating disorder	1	<1%
Other specified eating disorder	6	3%
Substance-Related and Addictive Disorders	11	6%
Personality Disorders	8	5%

Note: Diagnoses were based on a chart review and do not necessarily reflect patients' current diagnoses at the time of treatment. Percentages add up to over 100% due to comorbidity.

Table 2

Power Analyses

Statistical Analysis	Estimated Effect Sizes	Required Sample Size (n)
Aim 1		
Paired t-test	d = 0.62	66
McNemar's test	Odds ratio = 24 Prop discordant pairs = .08	67
Multiple linear regression: Interaction between HT performance and scoring criteria on outcome variables	f^2 = .10	81
Multiple linear regression: Association between HT performance and outcome measures for each scoring criteria	f^2 = .15	128
Aim 2		
Multiple linear regression with random intercepts: Main effect of verbosity on HT performance	β = .20	160
Multiple linear regression with random intercepts: Interaction between verbosity and scoring criteria on HT performance	β = .20	145

Note. HT = Hinting Task. All calculations are for power = 0.80

Table 3

Bivariate Correlations Between Measures

	Original HT	***Revised HT***	***Verbosity***	***RMET***	***PQ-B***	***PHQ-9***	***BASIS-24-R***
Revised HT	.73***						
Verbosity	.06	.23					
RMET	.25*	.22	-.01				
PQ-B	-.25*	-.18	.05	-.26*			
PHQ-9	-.08	-.04	-.09	.04	.35***		
BASIS-24-R	-.07	-.03	-.07	-.18	.41***	.41***	
Age	.00	-.02	-.14	-.10	-.10	-.19	-.02

Note: HT = Hinting Task; RMET = Reading the Mind in the Eyes Test (% correct); PQ-B = Prodromal Questionnaire-Brief; PHQ-9 = Patient Health Questionnaire-9; BASIS-24-R = Behavior and Symptom Identification Scale, Relationships Subscale (Average item response). *$p < .05$ **$p < .01$ ***$p < .001$

Table 4

Descriptive Statistics and Demographic Associations for Primary Measures

Variable		Age		Ethnicity			Sex				
	M (*SD*)	*r*	*p*	*F(1,171)*	η²	*p*	Female: *M* (*SD*)	Male: *M* (*SD*)	*t(172)*	*d*	*p*
Original HT	16.85 (2.43)	.00	.93	1.21	.01	.64	16.75	16.71	0.00	0.00	>.99
Revised HT	15.36 (2.23)	.02	.93	0.66	.00	.72	15.19	15.28	-0.08	-0.01	>.99
Verbosity	166.6 (93.61)	-.13	.26	0.42	.00	.72	170.41	190.79	-0.84	-0.15	.94
RMET	69.92 (10.94)	-.09	.33	2.97	.02	.61	71.03	67.29	2.05	0.34	.15
PQ-B	19.33 (4.09)	-.08	.33	1.41	.01	.64	5.08	5.05	0.33	0.05	>.99
PHQ-9	15.15 (6.10)	-.19	.08	0.00	.00	.99	16.08	13.03	3.04	0.51	.02
BASIS-24-R	1.42 (0.81)	.02	.93	0.12	.00	.86	1.44	1.44	-0.10	-0.02	>.99

Note: HT = Hinting Task; RMET = Reading the Mind in the Eyes Test (% correct); PQ-B = Prodromal Questionnaire-Brief; PHQ-9 = Patient Health Questionnaire-9; BASIS-24-R = Behavior and Symptom Identification Scale, Relationships Subscale (Average item response). *p*-values adjusted for multiple comparisons using false discovery rate procedure.

Table 5

Interaction and Simple Main Effects of Hinting Task Performance and Scoring Criteria on Outcome Measures

	Interaction: Scoring Criteria by Performance		Simple Main Effect: Original Criteria HT Scores		Simple Main Effect: Revised Criteria HT Scores	
	β	*p*	*β*	*p*	*β*	*p*
RMET	-.02	.83	.22	**.04**	.19	**.049**
PQ-B	.05	.68	-.18	.07	-.12	.22
PHQ-9	.04	.71	-.11	.22	-.06	.53
BASIS-24-R	.04	.71	-.05	.57	-.01	.91

Note: RMET = Reading the Mind in the Eyes Test (% correct); PQ-B = Prodromal Questionnaire-Brief; PHQ-9 = Patient Health Questionnaire-9; BASIS-24-R = Behavior and Symptom Identification Scale, Relationships Subscale (Average item response). Values determined by 12 separate regressions. *p*-values are adjusted for multiple comparisons using false discovery rate procedure. Age, sex, and ethnicity were included as covariates. Bolded values indicate statistical significance.

Figure 1

Example Hinting Task Item and Revised Scoring Criteria

10. Maria is getting off the train with three heavy suitcases. John is standing behind her. Maria says to John, "Gosh! These suitcases are a nuisance."

Question: What does Maria really mean when she says this?

_____

_____

Additional Prompt: Maria goes on to say, "I don't know if I can manage all three."
Question: What does Maria want John to do?

_____

_____

10. **Required components for a 2-point answer:**

- *Subject/verb* (She wants…; Can you…?)
- The idea that she wants *John/him specifically* to do it
- The concept of *helping **and/or** carrying the suitcases/them*

Common 2 point responses:
She wants him to help her with the suitcases
"Could you help me out here, please?"

Other Common responses:
Help her; Help carry them = give second hint or if this response is given on second hint, score 1 point
Take them; Take one of them = give second hint or if any of these responses are given on second hint, score 0 points

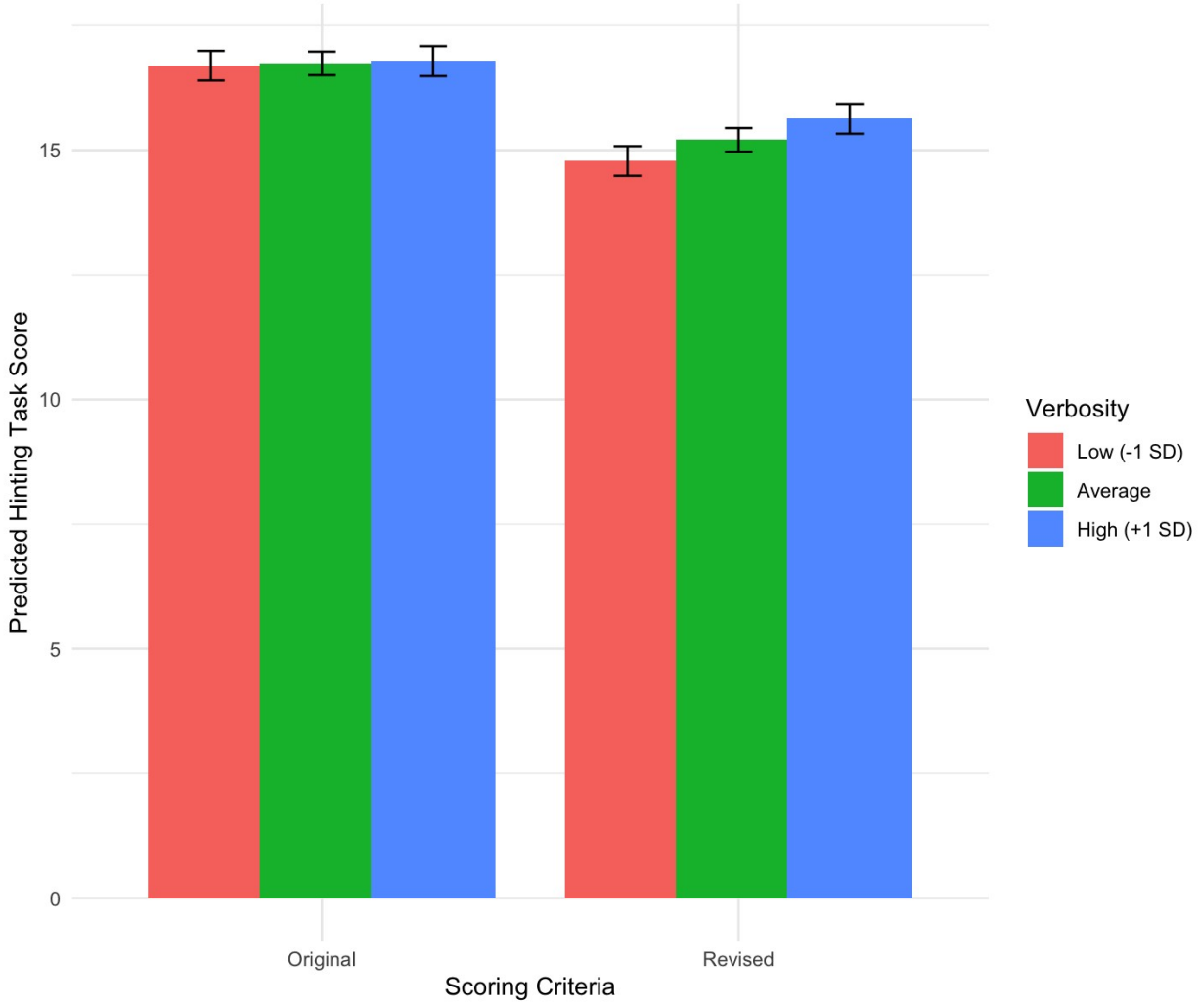
Figure 2

Comparison of Mean Hinting Task Performance Across Scoring Criteria

Note: The figure above depicts mean Hinting Task performance for original and revised scoring criteria. Data points represent individual participants. Horizontal lines indicate mean score. Blue and red curves represent the distribution of Hinting Task performance when applying the original and revised criteria, respectively.

Figure 3

The effect of scoring criteria on average Hinting Task scores at different levels of participant verbosity



Note. Error bars represent standard errors.