UC Davis
UC Davis Electronic Theses and Dissertations

Title

A Psychometric Perspective on Theory of Mind: Novel Methods and Measures for Studying Children's Social Understanding

Permalink

https://escholarship.org/uc/item/6zj9j6jt

Author

Heise, Megan Janelle

Publication Date

2023

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

A Psychometric Perspective on Theory of Mind: Novel Methods and Measures for Studying Children's Social Understanding

By

MEGAN JANELLE HEISE
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Lindsay C. Bowman, Chair

Simona Ghetti

Yuko Munakata

Committee in Charge

2023

Table of Contents

Abstract

The capacity to reason about and project mental states, such as intentions, desires, thoughts, knowledge and emotions, onto others' minds is fundamental to social cognition, and is the foundation of social behavior. A sophisticated understanding of the mind is referred to as a 'theory of mind' (ToM), and is characterized by an understanding of the mind as representational—meaning that it can reflect or hold information independent of the real world—and person-specific—meaning that each individual has their own mind and mental states. Thousands of studies have shed light on the developmental improvements in ToM across childhood, differences in ToM development across neurotypical and neurodivergent populations, and cognitive and environmental factors that support ToM development. However, there remain large unanswered questions that continue to puzzle the field regarding infants' seeming ability to pass ToM measures when assessed with non-declarative measures, identifying the social brain network and role of the temporo-parietal junction in ToM development, and how to characterize ToM more broadly as either a collection of discrepant mental states or a single social cognitive construct. However, empirically testing these questions is challenging due to limitations in ToM measurement tools. This dissertation illustrates the utility of psychometric examinations in further refining newly-developed ToM measures and in re-evaluating fundamental questions in ToM development, such as how mental states scale in difficulty across childhood.

Study 1 (n = 224) evaluates a multi-mental state measure of ToM developed for children ages 3- to 12-years-old (Richardson et al., 2018) within an item response theory framework in order to examine difficulty—how challenging items within the scale are—and discrimination—how well individual items reflect children's ToM. Results showed that items were generally discriminating, and the measure as a whole was reliable in assessing ToM in a more narrow

range of 3- to 7.5-years-old. Findings also expand the developmental trajectory of mental-state reasoning previously described by Wellman and Liu (2004), in which children reasoned about beliefs based on past experience, emotions, morals, and sarcasm generally after understanding false belief. However, this ToM scale included an over-representation of questions about belief reasoning, and did not include equal numbers of tasks from each mental state. Thus, in a second study we developed a novel measure of ToM that was designed with approximately equal numbers of items within each mental-state category.

Study 2 (n = 206) develops and evaluates a multi-assessment multi-mental state measure of ToM—the Comprehensive Assessment of Theory of mind (CAT)—in order to create a composite that reflects the range of mental-state reasoning encompassed by ToM, includes mental states not previously included in the Wellman and Liu (2004) scale (e.g., visual perspective taking), in addition to a non-social representational measure (e.g., false sign), and examines the developmental trajectory of different mental states when multiple assessments are administered. Results show that items were highly discriminating, and the measure reliably assessed ToM in 3- to 8-year-old children. Results also suggest a less-straightforward developmental trajectory of ToM when assessed with repeated measures and multiple question-types.

These two large-scale studies lay critical groundwork for further psychometric work in ToM research, and illustrate how the inclusion of additional mental states in ToM assessments leads to a richer understanding of how children reason about mental states. We encourage future extensions of these psychometric inquiries and of the Comprehensive Assessment of Theory of mind (CAT).

Acknowledgements

This dissertation reflects the amalgamation of my academic journey and personal curiosity about how we understand (and measure) the developing mind. I am so grateful to have had phenomenal mentors take me under their wings and shape the scientist that I am today.

Dr. Lindsay Bowman, my dissertation chair and mentor since 2013, has been instrumental in my academic growth, and I am grateful to have benefitted from her guidance these past ten years. Lindsay afforded me freedom to explore my interests to the fullest extent possible and I am lucky to have her in my corner. I am especially thankful to my dissertation committee members—Dr. Simona Ghetti and Dr. Yuko Munakata—and my quantitative committee members—Dr. Mijke Rhemtulla and Dr. Emilio Ferrer, whose thoughtful input was critical in refining these two studies to their current form. Additionally, I am thankful for guidance from Dr. Kristin Lagattuta, Dr. Lisa Oakes, Dr. Erie Boorman, and Dr. John Olichney, who served on various committees during my graduate studies and directed me towards diverse domains in psychology, enriching my understanding of the mind and brain.

I would not have applied to Ph.D. programs without the guidance of my undergraduate mentors, Dr. Stephanie Carlson and Dr. Megan Gunnar at ICD at the University of Minnesota. They inspired me with their research, were extraordinarily generous with their time, and envisioned my future in research before I knew this career was possible. Additionally, Dr. Angel Lillard at the University of Virginia fully involved me in the research process in my time as a laboratory coordinator, and encouraged some of my earliest research questions and ideas.

I would also like to thank my family, and friends who feel like family—especially Michaela DeBolt—who has been my other half these past six years. Finally, thank you to the hundreds of children and their caregivers who participated in these studies, without whom none of this research would have been possible.

Chapter 1: Psychometric issues in Theory of Mind research: General measurement problems, and analysis of a multi-mental-state theory of mind measure in 3- to 12-year-old children

Abstract

Theory of mind (ToM) encompasses diverse mental states and develops across the lifespan. The variability in constructs and methods related to ToM research has implications for measurement accuracy and reliability. Despite the need for psychometric evaluations, few studies have examined both scaling of individual tasks (i.e., difficulty) *and* how well individual items assess ToM (i.e., discrimination). We evaluated a newly developed measure of ToM that has 17 items and assesses 7 mental-state constructs (Richardson et al., 2018) and was administered to a large sample of typically developing children (M_{age} = 7.17 years, *SD* = 2.23, *n* = 224, 114 girls, 109 boys, *n* = 1 genderfluid trans-masculine child, 66% were White non-Hispanic/Latinx, and 34% identified as non-White or as Hispanic/Latinx). Results showed that items were highly discriminating, and the measure showed relational and construct validity. Analyses also suggest the measure may best capture ToM performance in children ages 3- to 7.5-years-old. Finally, results reveal a more complete 'scaling' of multiple mental state constructs including desires, beliefs, and false beliefs (as in Wellman & Liu, 2004) but also adds emotion understanding, moral reasoning, and interpretive ToM. We conclude with recommendations for psychometric analyses in our field and future use of this measure.

The past decade of research in developmental science has brought to light issues in measurement and methodology (See Nosek et al., 2022 for review). These issues are pervasive in the particular field of *theory of mind (ToM)* research, which involves investigation of the developing understanding of mental states (such as beliefs, desires and knowledge) as person-specific representations of the world that motivate behavior. The measurement of ToM represents a substantial portion of investigations across multiple fields in psychology, with lines of research into developmental (Carlson et al., 2013; Slaughter, 2015; Wellman et al., 2001), cognitive (Jara-Ettinger, 2019; Ong et al., 2018), social (see meta-analyses: Imuta et al., 2016; Slaughter et al., 2015; Schurz et al., 2021), clinical (Nilsson et al, 2016; Peterson et al., 2005), and neuroscientific (Bowman & Wellman, 2014; Grosse Wiesmann et al., 2017; Saxe & Powell, 2006) domains. While the breadth and variability in ToM research can be seen as a strength and testament to its far-reaching implications, it also creates risk for *psychometric issues.* Psychometric issues are problems with reliability, validity, and other factors that may limit a measure's ability to actually capture the target construct or developmental phenomenon. Psychometric analysis of ToM measures can illuminate these issues, and reveal the robustness and accuracy of existing measures for capturing stability, change, and individual differences in ToM. The present study highlights the importance and utility of using item response theory (IRT) models—a method of psychometric analysis commonly used in education research (e.g., Hori et al., 2020; Min & Aryadoust, 2021) but under-utilized in developmental science—to evaluate a multi-mental state measure of ToM development used in previous research (Richardson et al., 2018). In doing so, we confirm the utility of this measure for future research. More broadly, we showcase the value of psychometric analyses for psychological measures, using ToM as our specific example given the broad interest in ToM across multiple domains of

psychology, and given that the characteristics of ToM development and its measurement make it especially open to psychometric issues that should be evaluated and addressed.

The Need for Psychometric Analysis of ToM Measures

ToM is the understanding that people each have their own minds, that minds motivate observable behavior in the world (Wellman, 2014), and that our minds can either accurately or inaccurately reflect the real world (e.g., false belief, Wimmer & Perner, 1983). ToM encompasses a broad range of mental-state understanding, including understanding others' desires, knowledge, visual perspectives, and beliefs. ToM is also involved in making moral judgements (Killen & Smetana, 2013), and in understanding sarcasm or conversational tone (Happé, 1994; Peterson et al., 2012). Thus, ToM itself is a broad construct, and as such ToM measurement spans multiple ages (i.e., beginning in infancy and continuing throughout the lifespan; Henry et al., 2013; Wellman, 2018), modalities (e.g., looking time, verbal responses; see review: Sodian et al., 2020), and mental states (e.g., false belief, emotions, desires, knowledge; Wellman & Liu, 2004), which results in enormous measurement variability. This variability creates potential psychometric challenges.

To elaborate, considering the most widely studied aspect of ToM which involves administering behavioral tasks that require participants to make explicit verbal or pointing responses, there are two common practices in ToM measurement: one is to create a single 'composite' measure of ToM by combining performance across measures of multiple mental states (e.g., beliefs, desires, knowledge, emotion; Richardson et al., 2018; Shahaeian et al., 2011; Wellman et al., 2011), the other is to use a singular task or task type as representative of ToM more broadly (e.g., measure exclusively belief or false-belief reasoning, Devine & Hughes, 2014; Milligan et al., 2007; Wellman et al., 2001). However, each of these practices assumes that

ToM is a single unitary construct that can be assessed equally-well regardless of task (e.g., that both desires and beliefs reflect a singular ToM construct) and that measures of a single mental state can be used to draw conclusions about ToM more broadly (e.g., findings about false belief have implications for understanding ToM more broadly). Problematically, these assumptions are rarely empirically evaluated (though see Beaudoin et al., 2020; Quesque & Rossetti, 2020; and Warnell & Redcay, 2019 for some investigations into ToM measurement). Moreover, there is some evidence to suggest that performance across different mental-state measures is only weakly correlated (Warnell & Redcay, 2019), and that different mental state constructs (e.g., beliefs versus desires) are supported by different neural systems (Bowman et al., 2012; 2015). Thus, there is a lack of evidence to confirm that the assumptions guiding the field's measurement of ToM are accurate, and some evidence to suggest that assumptions may be wrong.

Psychometric analysis of ToM measures offers a way to empirically evaluate the appropriateness of a given ToM measure for capturing an intended developmental phenomenon. Common psychometrics include evaluations of *reliability* (measurement precision, Flora, 2020), *validity* (the extent to which a measure assesses the cognitive construct of interest, Cronbach & Meehl, 1955), *difficulty scaling of tasks* (standardized scores of how difficult or challenging an individual item is within a measure), and *discrimination of items* (how well an item differentiates between different ability levels). There may be reasons to be concerned about each of these psychometric properties when examining existing ToM measures.

Concerns with ToM Task Reliability

Reliability refers to the internal consistency of a measure, or how closely individual items are related to each other. Measures with high reliability have many items, have items that are highly related, and are used in samples with high variability across individuals (Flora, 2020).

ToM measures rarely achieve these standards. ToM measures typically include few items (e.g., one or two false-belief tasks, see discussion by Bloom & German, 2000), and if multiple mental states are evaluated there is often only a single task that represents performance on each mental state (e.g., Müller et al., 2012; Wellman et al., 2008). There is also limited variability in ToM performance depending on the age group selected. Current methods cannot distinguish whether limited variability in a task or measure is a result of an unreliable measure (i.e., a measure that does not accurately assess ToM) or a result of using the measure in the wrong age group (e.g., children are at ceiling performance). For example, there is little to no variability in behavioral performance on diverse desires – children's understanding that one can have difference desires/preferences than their own – measures by 5-years-old (Bowman et al., 2012; Henning et al., 2011), but this task is often included in broader batteries of ToM assessing children into middle childhood (Richardson et al., 2018). Including measures without variability may increase noise in the ToM measure because typically-developing 4-year-olds who fail diverse desires may fail due to inattention rather than a lack of ToM ability. Thus, better tools are necessary to identify which tasks produce variability in each age group (e.g., toddlers, preschoolers, adolescents, adults), and which combination of tasks can be used to develop a reliable battery across age ranges.

Furthermore, items do not always relate (e.g., measures of desires may not relate to measures of beliefs; see Warnell & Redcay, 2018). In a recent review of reliability, Beaudoin and colleagues (2020) found that reported reliability estimates ranged from $\alpha = 0.37\text{-}0.89$ (for emotion understanding) to $\alpha = .87\text{-}.91$ (for perspective taking). Moreover, only about 30% of studies on ToM reported reliability (e.g., internal consistency) of their selected ToM measure. These findings demonstrate an overall under-reporting of inter-item reliability in the ToM

literature, and variability in how reliable different ToM measures are, including a problematic range that includes poor and moderate reliability of ToM measures.

Low inter-item reliability in ToM measures will affect the ability to detect real effects within ToM research. For example, low reliability deflates simple correlations between variables (Metsämuuronen, 2022; Spearman, 1904). The impacts of low reliability are more complicated when fitting more complex statistical models, and can either deflate the relation between ToM and other variables, or *increase* the relation between ToM and other variables (e.g., in path models, in models in which there is low reliability in a covariate; See Nimon et al., 2012 for further discussion). Although there are ways to correct for attenuation due to low reliability (e.g., Metsämuuronen, 2022; Padilla & Veprinsky, 2012), this practice remains uncommon within ToM research. As noted above, the reliability of ToM measures tends to be lower than recommendations for good reliability of measures (e.g., reliability ranges of 0.37 [Iannotti, 1978] to 0.98 [Muris et al., 1999] when recommendations are .60 or higher for 'good' reliability; Cicchetti, 1994; Nunnally, 1978), suggesting that low reliability indices likely have far-reaching impacts for our understanding of the relations among mental states, and between ToM and other measures.

There are also issues in terms of how reliability analyses are conducted and reported in ToM research. Cohen's α, is the most commonly reported index of reliability (See Beaudoin et al., 2020), but this index may not be optimal for assessing reliability of ToM measures. To elaborate, Cohen's α, assumes that there is a unidimensional underlying factor (See Savalei & Reise, 2019). In other words, Cohen's α assumes that there is a single ToM factor that underlies all measures of mental-state reasoning in multi-mental state scales. However, this assumption is rarely tested (or at least results of such assumption tests are rarely reported), and some studies

assessing ToM have found multi-dimensional latent factors of ToM (See Altschuler et al., 2018; Osterhaus et al., 2016; Wang et al., 2021). More generally, the appropriateness of α as an accurate measure of reliability has been called into question given that psychological indices often do not meet assumptions of α; for example, that items are continuous and normally-distributed (See McNeish, 2018; Savalei & Reise, 2019).

One evaluation of reliability of ToM does show high agreement: inter-rater reliability in scoring children's responses (see Guajardo et al., 2009; Jin et al., 2017; Olineck & Poulin-Dubois, 2007). However, experimenter agreement does not necessarily give credence to the evaluation of the construct (e.g., a measure could be an unreliable assessment of ToM even if the scoring scheme is consistent). Further, many researchers report inter-rater reliability as the "percent agreement" across coders (over 20%, as described by Beaudoin et al., 2020). Percent agreement has been criticized for over 60 years (Cohen, 1960) as an over-estimation of reliability because it does not account for chance agreement (See Hallgren, 2012; McHugh, 2012), but remains widely used within the field (e.g., Choi & Luo, 2015; Hwa-Froelich et al., 2017; Mull & Evans, 2010; Olineck & Poulin-Dubois, 2007; O'Kearny et al., 2017; Phillips & Wellman, 2005). The interpretation of inter-rater reliability as being representative of scale reliability broadly is therefore flawed, and does not suggest good psychometric properties of the measure itself.

In sum, the status quo of ToM research is concerning in terms of how reliability of ToM measures is calculated, how often reliability analyses are reported, and the extent to which ToM measures meet acceptable reliability standards. New psychometric analyses are needed, as well as new ToM measures, in order to address these existing reliability issues in ToM research.

Concerns with ToM Task Validity

Validity refers to the ability of a measure to assess the underlying construct. There are multiple interpretations of validity, but some common ways of statistically assessing validity include examining correlations with related constructs (i.e., "relational validity"), examining relations with relevant constructs over time (i.e., "predictive validity", Cronbach & Meehl, 1955), or with other measures assessing the same construct (i.e., "construct validity", Cronbach & Meehl, 1955). There are issues concerning the validity of ToM measures in existing research, particularly in the relations between ToM and other constructs. Although there are general correlations between ToM and related constructs such as vocabulary and executive functioning, there are also inconsistencies and discrepancies in these correlations across development: Robust correlations between ToM and executive function exist in preschool (Devine & Hughes, 2014; Milligan et al., 2007), but are weaker and sometimes nonexistent in toddlerhood (Carlson et al., 2004, 2015), and findings in middle childhood are mixed (Devine et al., 2016; Lecce et al., 2017; Wilson et al., 2018). Similarly, correlations between ToM and other social constructs are also not consistent or straightforward (e.g., ToM is predictive of both antisocial—e.g., bullying, Fink et al., 2020; Stellwagen & Kerig, 2013—and prosocial behaviors—e.g., helping, sharing, Kuhnert et al., 2017). Of greater concern, performance on tasks assessing different mental state constructs (e.g., beliefs versus desires) are only weakly correlated and in some cases are uncorrelated in both childhood and adulthood (Warnell & Redcay, 2019). These conflicting and changing relations between ToM and relevant constructs may be due to fundamental aspects of ToM and its development, but they also raise questions about the validity of the measures used to assess ToM. Relatedly, there are open questions about the source of these conflicting and changing relations, which could be caused by a range of factors such as low powered samples, different relations with different target mental-state constructs, or confounds with age and task type

(Colquhoun, 2014, 2017; Halsey et al., 2015). Thus, psychometric analysis of large-scale studies that include multiple mental states and wide age ranges are needed to help shed light on potential issues of ToM measurement validity.

Rare Assessment of Item Difficulty and Discrimination

Beyond reliability and validity, there are other psychometric properties that are beneficial to examine in ToM measures. Specifically, there are open questions about the *difficulty* and *discrimination* of different items within ToM measures. Difficulty and discrimination are psychometric properties extracted from 2-parameter logistic (2PL) *item response theory (IRT) models* (Embretson & Reise, 2000), a family of latent factor models commonly used to assess test properties. In a 2PL model, children's ToM is estimated as a unidimensional latent factor, meaning that ToM is estimated as a single cognitive construct. The parameters in the 2PL model are item difficulty (how much of the latent trait ToM is needed to pass the item), and item discrimination (how well the item discriminates or differentiates between participants with different ToM abilities), see Formula 1.

Formula 1. Probability of a correct response in the 2PL IRT model.

$$P(X = 1|\theta, a, b) = \frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}$$

As illustrated in formula 1, in the 2PL model, the probability of a participant's correct response is dependent on the participant's underlying ToM (θ, the standardized latent factor), the item's discrimination (a, the slope showing how closely the item is related to the latent factor), and the item's difficulty (b, the standardized latent trait score when there is a probability of .5 of passing the item). Examination of both item difficulty and item discrimination is rare in ToM

research, although each has important merit for understanding ToM specifically (and psychological processes generally).

Difficulty. Difficulty refers to how much of a construct (e.g., ToM) is needed in order to pass a task. Evaluating the difficulty of various task items enables investigation of both how and when ToM develops. Despite the potential critical utility of evaluating task difficulty, few studies do so. Notable exceptions include difficulty scaling of the 'Theory of Mind Scale' developed by Wellman and Liu (2004) and used in subsequent studies across cultures and populations (Peterson et al., 2005; Peterson & Wellman, 2009; Shahaeian et al., 2011). These prior IRT models have been foundational in revealing clear patterns of children's ToM development. For example, item difficulty analysis revealed that children have more difficulty reasoning about belief-understanding compared to desire-understanding, and this consistent difficulty scaling occurs across cultures (Shahaeian et al., 2011) and in both typically and atypically developing preschoolers (Peterson et al., 2005). In contrast, the relative difficulty of belief- versus knowledge-reasoning changes depending on the culture tested (Wellman et al., 2006). Research has also extended the difficulty scaling of ToM items beyond preschool: Osterhaus and colleagues (2016; 2022) used Rasch models to examine difficulty scaling in advanced measures of ToM, including 2nd and 3rd order false belief and moral ToM longitudinally from 4- to 6-years-old, and in a sample of school-aged children 5- to 10-years-old. These studies demonstrated that more advanced measures of ToM are supported by different domain-general abilities in middle childhood compared to the traditional measures utilized in preschool.

Thus, the use of psychometric analyses to model item difficulty has informed the field's understanding of the developmental trajectories of different aspects of mental state

understanding from toddlerhood through middle childhood. Given the value of psychometric analysis of item difficulty, more of this type of analysis is needed, especially in tasks that include mental-state constructs beyond those originally included in the Wellman & Liu scales (e.g., visual perspective-taking, sarcasm, moral reasoning).

Discrimination. Another psychometric index that is possible in 2PL IRT analysis (but not possible in the prior described IRT analyses of Guttman scaling and Rasch models) is *discrimination*, which describes how well items within a scale differentiate between children with different abilities (e.g., different ToM performance). Discrimination can also be described as the strength of the relation between the item and latent factor (Reeve & Fayers, 2005). Therefore, items with high discrimination strongly relate to children's underlying ToM and are therefore reliable indices of ToM. As such, discrimination allows researchers to examine which ToM tasks most reliably assess ToM performance, and to select tasks that discriminate well within a target age range (e.g., toddlerhood, preschool, middle childhood). In extant ToM research, 'discrimination,' when it is rarely included in analyses, has been estimated as the correlation between a single item in the scale and all other items (similar to item-total calculations, e.g., Osterhaus et al., 2016). However, this calculation better reflects the internal consistency of a scale, rather than the discrimination of individual items. In contrast, no study of child ToM measures to our knowledge has evaluated the discrimination parameter within IRT models. Discrimination within the 2-parameter logistic (2PL) model describes the extent to which individual items discriminate between participants with high and low levels of ToM, after accounting for the item's difficulty. Thus, discrimination allows researchers to identify which items best measure ToM at different ToM abilities. This kind of psychometric analysis is particularly powerful for our understanding of ToM development because research continues to

develop new tasks and measures to best assess children's ToM, but there remain few empirical explorations of task reliability given constraints in testing children (e.g., cost of repeated second testing sessions, concerns with practice effects, child fatigue in completing long batteries). The discrimination parameter within the 2PL model does not require administering additional tasks or repeated testing sessions, and therefore offers a valuable approach to evaluating individual ToM tasks. However, discrimination has not commonly been evaluated in ToM research. Therefore, scales likely contain both 'good' and 'bad' ToM tasks, and the inclusion of bad tasks make the scale generally less valid and less reliable (because poor measures have high measurement error). Accounting for item-level characteristics such as item discrimination, as is done in the 2-PL IRT model, allows researchers to better isolate how well an item measures ToM, and thus is a much-needed approach.

Test Information Function. The *test information function* (TIF) describes which levels or 'amounts' of ToM the test most accurately measures. For example, a battery of challenging ToM measures including moral ToM, sarcasm and faux pas would be highly informative about the ToM ability of 7- and 8-year-olds, but likely not informative for 3-year-olds because these tasks do not discriminate well across ToM abilities in preschool. The TIF is valuable in addressing two common challenges in ToM research. First, the TIF allows researchers to empirically examine whether the targeted age range of the measure is correct. Specifically, researchers can examine whether the participants with ToM ability that is best measured by the task are within the expected age range for the measure. Second, there has been a shift in the field toward using composites of mental states in order to better capture broad ToM beyond belief reasoning. However, it is unclear whether the addition of more tasks results in a ToM measure that is also broad, or whether the measure as a whole still fundamentally is capturing a single

highly-discriminating mental state (e.g., belief reasoning). This question is also important because many of these measures are imbalanced, and there is a greater proportion of items assessing belief reasoning compared to other mental states (e.g., Bowman et al., 2017; Richardson et al., 2018). Although difficulty and discrimination are also valuable in identify reliable measures of ToM, the TIF is essential in understanding in which samples and ages to use ToM measures.

Summary of Psychometric Issues in ToM Research

Broadly, ToM research stands to benefit greatly from a psychometric re-evaluation of existing measures in our field, particularly of multi-mental state measures. Although prior psychometric evaluations have shed light on how mental-state understanding unfolds (e.g., the scaling of task order across countries, and special and typically-developing populations), there remain open questions about the overall reliability of measures (see Beaudoin et al., 2020 for review) and challenges with measurement validity across wide age ranges.

The Present Study: IRT Analysis of a Multi-mental State Measure of ToM in Preschool and School-age Children

The present study uses Item Response Theory (IRT) to evaluate the reliability, difficulty, and discrimination of a multi-item, multi-mental state ToM measure in a large-scale sample of typically developing children ages 3- to 12-years-old. We also assess validity by examining relations between children's ToM performance and performance on related constructs of executive functioning and language, and explore relations between our behavioral measure and parent-report measures of children's ToM. In doing so, we begin to address several gaps in our knowledge of the health of measurement of ToM development, and we showcase the utility of psychometric analyses more broadly.

We chose to psychometrically evaluate a recently developed ToM measure that consists of multiple story-based/vignette-style items to assess multiple mental-state constructs including common and diverse desires, diverse beliefs, false beliefs, beliefs based on past experiences, true beliefs, interpretive ToM, visual perspective taking, moral reasoning, emotions and sarcasm (Richardson et al., 2018). We targeted this particular measure given its potential to be highly impactful in ToM research: It has been used with an exceptionally large age range across both preschool and school aged children (whereas most measures are used with just one of these age groups alone), it is designed to capture both individual and group-level differences, and it has produced findings in behavioral and neuroscience domains (Richardson & Saxe, 2020),

Moreover, the assessment of this particular measure is especially useful given it is representative of the kinds of measures that are now being more commonly used in research on ToM development in preschool and school-aged children. Specifically, there is an increasing number of studies that have shifted away from using a single prototypical measure of ToM (e.g., a false-belief task) to alternatively using 'scales' that consist of several individual tasks that collectively test multiple mental-state understandings such as desires, knowledge, and beliefs (e.g., Hanson et al., 2014; Richardson et al., 2018; Shahaeian et al., 2011; Wellman et al., 2011). Psychometric analysis of these kinds of 'scale' or 'composite' measures (such as the one analyzed in the present study) is especially important. On the surface, using a measure that assesses more mental states is advantageous because it could more effectively measure variability in ToM, including individual differences across participants and age groups, and more closely reflects the broad and multifaceted nature of ToM. However, as noted above, these multi-mental-state scales also face challenges psychometrically: Items that assess multiple different mental states will be less consistent with each other within a scale compared to items that assess

a single mental state (e.g., false belief) (Dunn et al., 2014), decreasing the reliability and consistency of the overall composite measure. Given the increasing use of multi-mental-state composites and their potential advantages for studying ToM development, it is critical to identify measures that offer reliable and valid assessments of children's ToM. The present study adopts this aim.

Our use of a 2PL model as our particular model for psychometric analysis offers important advances over existing psychometric analyses in ToM research. In contrast to prior work, we examine item discrimination in addition to difficulty through the 2PL IRT model in order to determine whether individual items within the scale are accurate measures of children's underlying ToM. Finally, our use of the TIF identifies which ages the multi-mental-state ToM measure best assesses, and whether children's performance in this composite reflects a diversity of ToM reasoning or only select tasks that show variability or that are over-weighted in the composite (e.g., belief reasoning). Thus, our selected analysis model offers distinct advantages over prior work using Guttman scaling, Rasch analyses, or Cronbach's α alone (Osterhaus et al. 2016; 2022; Shahaeian et al., 2011; Wellman & Liu, 2005;).

Indeed, results from our psychometric analysis of this multi-mental-state measure will have several implications for advancing ToM research. First, our psychometric analysis can shed light on how different aspects of mental-state understanding build and progress through childhood by examining difficulty scaling of ToM measures (as was done in Wellman & Liu, 2004). Although there have been decades of research documenting how mental state understanding develops throughout preschool, there remains more limited research in how ToM develops beyond an understanding of false belief. For example, there are several mental states that emerge beyond age 5, including a deeper understanding of emotions and tone through

sarcasm (Happé, 1994), and the conditions under which two people would have the same or different interpretations of the same image (referred to as 'interpretive ToM', see Taylor et al., 1991; Carpendale & Chandler, 1996; Lagattuta et al., 2010). However, there is minimal work investigating how to scale these more advanced aspects of mental state understanding, and which of these mental states are the most challenging to acquire. Thus, examining the difficulty parameter in the 2PL model of our target ToM measure that includes items assessing these more advanced aspects of ToM will yield a more complete perspective on the developmental trajectory of ToM through preschool and middle childhood.

Second, this study will determine the most accurate items to assess ToM within the composite using the discrimination parameter within the 2PL model. Examining item discrimination will distinguish which mental states should be included as a reliable measure of ToM within a composite. Developing robust and reliable composites is critical to ensure that variability in ToM scores is due to underlying ToM ability and not to low reliability of the ToM measure, and examining discrimination is necessary to increase replicability of findings and to ensure that children's performance on tasks accurately reflect their underlying cognition. Thus, we will use estimates of discrimination in combination with difficulty estimates in order to evaluate our multi-mental-state ToM measure, and inform the development of a more accurate composite of items.

Third, we will examine the reliability of the selected subset of ToM tasks within the measure by evaluating both internal consistency (i.e., with Cronbach's α) and how well the scale performs across different ability levels (i.e., with the *test information function, TIF)*. The TIF is an additional assessment of reliability that is extracted from the IRT model and offers insight into not only whether items are related, but also what levels of ToM understanding the measure

is best suited for. The TIF therefore offers a more complete picture of the range of ToM abilities (e.g., low, moderate, or high performance) and the corresponding items the ToM measure can best reveal, beyond examining how closely items are correlated to each other by calculating Cronbach's α alone. We will also assess both relational validity through examining correlations between ToM and related domain-general constructs (e.g., executive function, vocabulary), and construct validity through examining correlations between our multi-mental-state measure and two established questionnaire-format parent-report measures of ToM (CSUS, Tahiroglu et al., 2014; TOMI-2, Hutchins et al., 2014).

Finally, our psychometric analysis can also shed additional light on how we conceptualize of ToM theoretically as a construct. Most existing research in ToM assumes that ToM is a single cognitive construct that can be assessed by examining children's performance on different aspects of mental-state understanding (e.g., desires, beliefs, sarcasm), as indicated by the creation of sum scores and our field's choice of Cronbach's α as a measure of internal consistency—each of these practices statistically assume that all items are measuring a single underlying ToM ability. Few studies have examined the factor structure of the storybook-style tasks that are most commonly used in the field in the preschool age range. Therefore, our assessment of the factor structure of this task (as required in testing assumptions for the 2PL IRT analysis) will take a first step in examining the factor structure of a wide-range of mental-state items that are matched in task format.

Overall, our 2PL IRT analysis of our targeted multi-item multi-mental-state measure, coupled with our assessment of both relational and construct validity, can reveal a specific measure that can be used confidently in future ToM research. In particular, we evaluate a ToM measure designed for a wide age range and that encompasses a broad array of mental states, and

thus has potential to be widely used in the field. In addition to this critical evaluation of an existing ToM measure, we also demonstrate the broader utility of psychometric analyses of ToM measures more generally, and our methods and approach here can be adopted in future research.

Method

A total of 226 children were administered the "Theory of Mind Booklet 2"—a measure of explicit ToM (Richardson et al., 2018). Data were compiled from two laboratories: 127 children were tested from the greater Boston area before the onset of COVID-19, and 99 children were tested from the Sacramento Valley (n = 33 in a laboratory setting, n = 66 remotely during the COVID-19 pandemic). All procedures were approved by the Institutional Review Boards across both institutions and parents provided informed consent prior to their child's participation. Families in Sacramento were compensated with a $10 giftcard for participating. Families in Boston were compensated with a $50 giftcard for participating, as all children completed an fMRI scan in addition to the behavioral ToM task.

Participants

The final sample of participants included in analyses consisted of 224 children ages 3- to 13-years-old (M_{age} = 7.17 years, SD = 2.23, n = 114 girls, n = 109 boys, and n = 1 genderfluid, trans-masculine child), after excluding 2 participants from the Sacramento sample who were being screened for autism.

Participants' ethnic and racial backgrounds were representative of the community from which they were recruited. Of children tested in the Sacramento Valley, 63 children were White (22.22% Mexican, Chicano/a, Puerto Rican, Latinx, or Spanish-origin), 24 children were multi-racial (41.67% Mexican, Chicano/a, Puerto Rican, Latinx, or Spanish-origin), 5 were Asian (20% Mexican, Chicano/a, Puerto Rican, Latinx, or Spanish-origin), 3 children were Black (33.33%

Mexican, Chicano/a, Puerto Rican, Latinx, or Spanish-origin), and 2 parents chose not to disclose. Of children tested in the greater Boston area, 81 were White (3.75% Mexican, Chicano/a, Puerto Rican, Latinx, or Spanish-origin), 10 were multi-racial, 5 were Asian, 1 was Black, and 30 parents chose not to disclose.

Measures

Focal Theory of Mind Measure

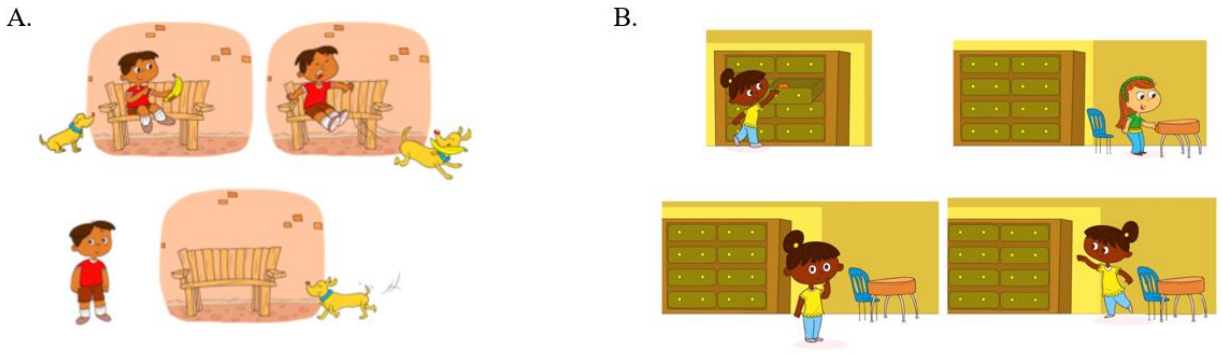
Theory of Mind Booklet 2 (Richardson et al., 2018). ToM was assessed by an interactive story booklet previously used with children ages 3- to 12-years-old. It includes a broad range of ToM tasks targeting common and diverse desires, diverse beliefs, false beliefs, beliefs based on past experiences, true beliefs, interpretive ToM, visual perspective taking, moral reasoning, emotions, and sarcasm. Participants tested in lab were administered the task in person using a picture booklet with experimenter and child seated together at a table. Participants tested in person could respond to test questions verbally or by pointing. Participants tested remotely were administered an animated presentation consisting of digital versions of the exact same booklet stimuli delivered through PowerPoint with experimenter and child seated at a computer and interacting over Zoom. Participants tested remotely responded verbally. As described below, although the administration of the task varied across different portions of the sample (e.g., in-person Boston environment, in-person Sacramento environment, online environment), there were only a few differences in task performance across testing environments that could each reasonably be attributable to expected age-related and SES-related differences as has been demonstrated in prior research (Ebert et al., 2017; Pluck et al., 2021; Wellman et al., 2011). As described in Appendix S1, the protocol for online testing ensured proper comprehension of the

task, and is in line with interactive online ToM testing protocols that were found to be empirically equivalent to in-person testing (Bambha & Casasola, 2021; Shidelko et al., 2021).

The ToM Booklet 2 consists of eighteen unique stories, with 57 items total across all stories. Stories involved one to three central story characters who acted out various familiar daily activities (e.g., packed a lunch with their parent, stored a snack in their school cubby). After hearing the story, participants were asked a series of questions about the characters' mental states given the context of the events in the story. See Figure 1 for examples, and https://osf.io/spqgc/ for full measure. There were three types of items across the stories: control items that assessed story comprehension, prediction items in which the child was asked to predict a character's actions based on the character's mental state or to predict how a character would feel, and explanation items in which the child was asked to give their reasoning for their stated prediction.

Figure 1

Sample Stories from the ToM Booklet 2



Notes. Panel A: Participants were told that they are viewing a picture from last week (top two images), and that last week a puppy stole Bill's snack. Today, as Bill gets his snack out, he sees the puppy coming toward him (bottom image). Participants are asked how Bill will feel when he sees the puppy – happy or sad? Panel B: Diana puts her snack in the drawer (top left). While

Diana is outside playing, someone moves her snack from the drawer to the desk (top right). Participants were asked where Diana will look for her snack (bottom left). After participants predict Diana's action, the final image is revealed (bottom right) and participants were asked to explain why Diana will search in the drawer.

Coding schemes for explanation items differed across Boston and Sacramento area testing sites. Thus, to achieve consistency, only prediction items were analyzed. To meet assumptions for item response theory (e.g., local independence, see Results), if stories contained more than one prediction item, the first prediction item from each story was analyzed. Thus, there were a total of 17 prediction items across the 18 stories (one story only contained an explanation question), with 2 prediction items about desires (one for common desires and one for diverse desires), 7 prediction items about beliefs (1 about diverse beliefs, 3 about false beliefs, 2 about beliefs based on past experiences, and 1 about true beliefs), 2 about visual perspective taking, 2 about moral reasoning, 2 about emotions, 1 about interpretive ToM, and 1 about sarcasm. For lengthy stories that required following a detailed set up, participants were required to answer a control/memory check question that assessed story comprehension. Five of the 17 stories included such control items after the child's prediction. Following scoring guidelines by Bowman and colleagues (2017) and Wellman and Liu (2004), if participants failed a control/memory check item that indexed story comprehension, they were scored as failing the accompanying prediction item unless the script indicated that experimenters should correct the participant's response. This scoring protocol reduces the potential that some correct responses represent "lucky guesses". Two additional items (#15.2 and #16.5) contained control items in which the participant was corrected before they answered the prediction item, and therefore these

control items were not considered when scoring the participant's prediction. See Table 1 for a brief description of the contents of each prediction item analyzed.

Table 1. Items Analyzed from Richardson and colleagues' (2018) ToM Booklet 2 and Scoring Criteria

Item number	Item description from Richardson et al.	Mental state category	Item description
1.1	Common desires	Common desires	Participants were told Henry preferred the same snack as the participant. Participants made a correct prediction if they identified Henry's snack preference.
2.1	Diverse desires	Diverse desires	Participants were told Kari preferred a different snack than the participant. Participants made a correct prediction if they identified Kari's snack preference.
3.1	Diverse beliefs	Diverse beliefs	Participants were told that Joe's snack could be in his backpack or in his desk. After guessing where the snack is, participants were told Joe thinks his snack is in the other location. Participants made a correct prediction if they identified Joe's belief about his snack's location.
4.1	Reference: easy	Visual perspective taking	There is a box of raisins in front of and behind Carmen. Carmen says, "I want *that* snack." Participants made a correct prediction if they said that Carmen wants the raisins in front of her.
5.1	False belief: reality known	False belief	Diana puts her snack inside a drawer. While Diana is outside playing, a girl moves Diana's snack to inside the desk. Participants made a correct prediction if they said that Diana would look for her snack in accordance with her false belief (inside the drawer).
7.1	False belief	False belief	Allie places her snack bag on top of a shelf. While Allie is outside playing, her snack bag falls behind the shelf. Sue places an identical-looking snack bag on top of the shelf. Participants made a correct prediction if they said that Allie would search on

			top of the shelf in accordance with her false belief (rather than behind the shelf).
8.2	Expectations	Belief based on past experience	Luke's dad has packed him crackers for snack everyday this month. Today, Luke's dad runs out of crackers and packs chips instead. Participants made a correct prediction if they said that Luke will think that there are crackers in his snack bag.
10.1	Reference: hard	Belief based on past experience	Josh tried celery for the first time today, and he liked it. Later, mom was packing Josh's snack for lunch, and mom asks Josh which snack he would like. There is celery near the glasses (drinking glasses), and carrots near the glasses (eye glasses). Josh says that he would like the one near the 'glasses.' Participants made a correct prediction if they said that mom would pack Josh carrots because she thinks he is referring to the 'glasses' by the carrots because has she does not know about his recent experience with celery.
11.2	Sarcasm control	Emotion	Charlie tells his dad that he wants a salty snack. Charlie's dad packs him popcorn. When Charlie sees the popcorn, he says, "Mmm, my favorite!" Participants made a correct prediction if they said that Charlie feels happy that his dad packed popcorn.
12.1	Reference: hard	Visual perspective taking	There are a small, medium, and large cookie in the kitchen. Marie can see all three cookies, but the large cookie is under a shelf and Marie's dad can see only the small and medium cookie. Marie's dad asks which cookie she would like, and Marie says that she would like the big cookie. Participants made a correct prediction if they said that dad would pack the medium-sized cookie which is the biggest cookie he can see.
13.1	Interpretation	Interpretative ToM	Melanie is drawing a picture so that her teacher knows which snack is hers. On the shelf there are two round fruits: an apple and an orange, and Melanie brought the orange for snack.

			Melanie draws two pictures in black and white: a picture of an orange slice (semi-circle with clear orange 'sections'), and a round circle representing the whole orange. Participants made a correct prediction if they said that Melanie should give her teacher the orange slice picture because he could think that the picture of the whole orange is a picture of an apple even though Melanie intended for it to depict an orange.
14.1	Emotion reminder	Emotion	Last week, Bill took his snack outside and a puppy stole Bill's snack and ran away with it. Today, Bill decides to eat his snack outside again. He sees the puppy coming toward him. Participants made a correct prediction if they said that Bill would feel sad when he sees the puppy.
15.2	True belief	True belief	Jenni put her snack on a shelf, but her teacher wanted to clean the shelf. Her teacher tells the class that he is moving all of the snacks to the desk. Participants made a correct prediction if they said that Jenni would search on the desk for her snack because she heard the teacher's announcement.
16.5	Moral reasoning	Morals	Melanie steals Mitch's cookie from his desk while he is not looking, and Melanie eats it. Participants made a correct prediction if they said that Melanie was being mean and naughty for stealing the cookie.
17.1	Morals/deception	Morals	The teacher brings cookies in for his class, and there are enough cookies for each student to have only one. Brian takes two cookies and places the extra cookie in his desk. Brian tells his teacher that there are no cookies left. Participants made a correct prediction if they said that Brian was lying to his teacher because there was a cookie in his desk.
18.1	Second-order false belief	False belief	Emma tells her dad that she would like string cheese for snack. While she is upstairs, she changes her mind and tells her dad that

			she would like yogurt instead, but her dad does not hear her. He goes to pack string cheese, but they are out of string cheese so he packs yogurt for Emma's lunch. Dad tells Emma that he couldn't pack the snack that she wanted. Participants made a correct prediction if they said that Emma's dad will think that Emma wants string cheese (first-order false-belief prediction item).
19.2	Sarcasm	Sarcasm	Leslie tells her mom that she would like something sweet for snack. Leslie's mom thinks that Leslie should eat something healthy so she packs carrot sticks. Participants made a correct predication if they say that Leslie actually is upset that her mom packed carrot sticks even though she said "Mm, my favorite."

Notes. Control questions were included for items 5.1, 7.1, 12.1, 17.1 and 18.1. See Richardson et al. (2018) and https://osf.io/grn76/ for complete measure.

Additional measures

Several additional measures were administered including standard parent-report ToM (Theory of Mind Invetory-2, Hutchins et al., 2014; Children's Social Understanding Scale, Tahiroglu et al., 2014), and standard measures of executive functioning (Up/Down, Kramer et al., 2015 and Happy/Sad, Lagattuta et al., 2011 in the Sacramento sample; Dimensional Change Card Sort, Zelazo, 2006 in the Boston sample) and vocabulary (Kaufman Brief Intelligence Test 2, Kaufman & Kaufman, 2004 in the Sacramento sample; Peabody Picture Vocabulary Test 4, Dunn & Dunn, 2007). In analyses that included these additional measures, only the subsamples that received them (as outlined above) were analyzed. See Appendix S2 for more detailed descriptions of each of these additional measures.

Results

All analyses were conducted in R (Version 3.6.2; R Core Team, 2019), in which Cronbach's α was calculated in the sjPlot package (Version 2.8.14; Lüdecke, 2023), exploratory factor analyses were conducted using the *scree* and *fa.parallel* functions in the psych package (Version 1.8.12; Revelle, 2019), confirmatory factor analysis was conducted in the lavaan package (Version 0.6-11; Rosseel, 2012), and IRT analyses were conducted in the ltm package (Version 1.1-1; Rizopoulos, 2006) and mirt package (Version 1.37.1; Chalmers, 2012).

We first examined whether there were meaningful differences by geographic location (Sacramento or Boston area) and by testing environment (in a laboratory or remotely), and we evaluated missingness in each of the 17 items in the measure. Details of these preliminary analyses are found in Appendix S3. In sum, participants tested in the Boston sample and in the online remote Sacramento testing sessions were slightly older ($M_{\text{ageBoston}}$ = 7.56, *SD* = 2.75; $M_{\text{ageSacramentoRemote}}$ = 7.02, *SD* = 0.92; $M_{\text{ageSacramentoLab}}$ = 5.97, *SD* = 1.09) but ToM performance

did not significantly differ across groups. Missingness was minimal (11 items had no missing data, and remaining items had 1 – 33% missing) and was handled using maximum likelihood for confirmatory factor and IRT analyses.

We then conducted our focal analyses as described in the sections that follow. We began by assessing item difficulty and discrimination using the 2PL IRT Model (described below). We used the results of this psychometric analysis to adjust the focal ToM booklet measure (e.g., remove certain items). We then evaluated the reliability and validity of this adjusted measure.

Assessing Item Difficulty and Discrimination in a 2PL IRT Model

Meeting IRT Model Assumptions

We first checked that data met assumptions for item response theory. Specifically, IRT models need to meet two focal assumptions: that items are locally independent (i.e., items share variance through the latent variable and are otherwise independent of each other), and that the latent factor is unidimensional (i.e., it is best represented by a single underlying latent variable). Exploratory and confirmatory factor analyses were used to ensure that data met assumptions of unidimensionality. All assumptions were met, see Appendix S4 for details.

In an effort to best represent the original measure by Richardson and colleagues (2018), to assess the maximum number of items in the IRT framework, and because model fit was acceptable, all 17 items were retained in IRT analyses in order to best evaluate the ToM measure's psychometric properties. Item descriptions are provided in Table 1.

Assessing Item Difficulty: 'Scaling' Multiple Mental-states

To assess item difficulty, we examined difficulty parameter estimates extracted from the item characteristic curve for each item, see Table 2 and Figure 2. We evaluated IRT parameters

such that a ‘good’ measure should have *some items with low difficulty* and *some items with high difficulty* in order to capture variability in ToM understanding where it exists across participants.

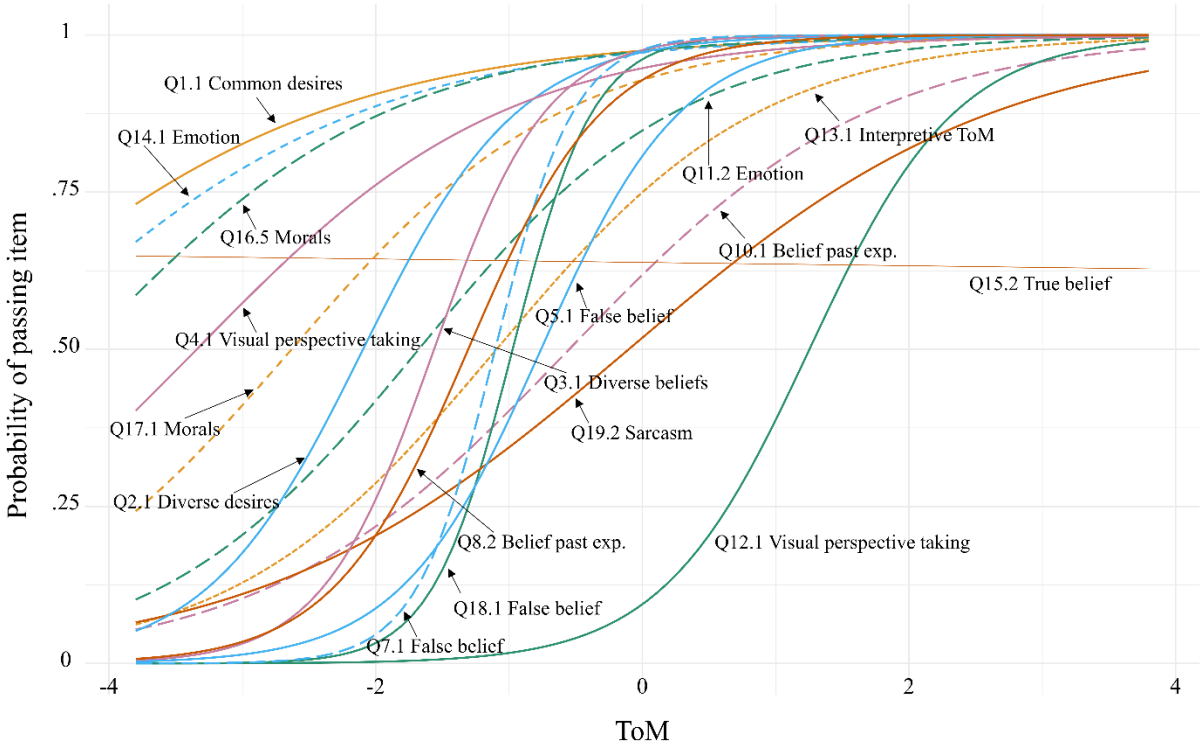
Table 2. Difficulty and Discrimination Estimates from Item Characteristic Curves

Item number	Mental state category	*N*	*M* (*SD*)	*a*	*b*
1.1	Common desires	224	0.97 (0.17)	0.70 (0.46)	-5.22 (3.04)
2.1	Diverse desires	224	0.92 (0.27)	1.70 (0.48)	-2.09 (0.35)
3.1	Diverse beliefs	224	0.89 (0.32)	2.37 (0.61)	-1.56 (0.19)
4.1	Visual perspective taking	224	0.93 (0.26)	0.86 (0.34)	-3.34 (1.07)
5.1	False belief	151	0.75 (0.43)	1.89 (0.48)	-0.76 (0.17)
7.1	False belief	224	0.82 (0.38)	3.38 (0.81)	-1.10 (0.12)
8.2	Belief based on past experience	218	0.83 (0.38)	1.96 (0.42)	-1.31 (0.18)
10.1	Belief based on past experience	224	0.60 (0.49)	0.88 (0.21)	-0.55 (0.20)
11.2	Emotion	187	0.84 (0.37)	1.03 (0.34)	-1.68 (0.51)
12.1	Visual perspective taking	189	0.21 (0.41)	1.81 (0.57)	1.25 (0.23)
13.1	Interpretive ToM	223	0.71 (0.46)	1.00 (0.23)	-1.09 (0.24)
14.1	Emotion	222	0.96 (0.19)	0.75 (0.44)	-4.75 (2.43)
15.2	True belief	224	0.64 (0.48)	-0.01 (0.16)	48.61 (676.26)
16.5	Morals	223	0.96 (0.19)	0.87 (0.44)	-4.20 (1.81)
17.1	Morals	189	0.92 (0.27)	0.97 (0.42)	-2.63 (1.01)
18.1	False belief	216	0.80 (0.40)	3.31 (0.77)	-0.97 (0.12)
19.2	Sarcasm	189	0.54 (0.50)	0.72 (0.24)	-0.10 (0.24)

Notes. Sample size (*N*), means and standard deviations, and item discrimination (*a*) and difficulty (*b*) with standard errors in parentheses.

Figure 2

Item Characteristic Curves from the 2PL IRT Model



Notes. Item characteristic curves illustrate the probability of passing each item (y-axis) as a function of item difficulty (*b*, illustrated by the line's position along the x-axis), item discrimination (*a*, illustrated by the slope of the line's curve, in which more discriminating items show steeper curves), and child's standardized latent ToM ability (θ, on the x-axis). Different colors and line styles (dashes, solid) are to aid in discriminating ICC curves visually.

Difficulty estimates (*b*) report the amount of ToM necessary (as a level of the latent factor ToM) to have a probability of .5 of passing an item. Therefore, items with positive

difficulty mean that items were more challenging (i.e., a higher-than-average ToM is needed to pass the item), whereas items with negative difficulty mean that items were easy (i.e., a lower-than-average level of ToM was needed to pass the item). Difficulty is valuable in uncovering the order in which mental states develop, and therefore we examined this parameter to examine the scaling of different mental states.

In examining item scaling in the difficulty parameter, the present study replicated prior research showing children can pass tasks assessing understanding of desires (items #1.1 and #2.1), before passing diverse beliefs (item #3.1), then false belief tasks (items #5.1, #7.1 and #18.1), then sarcasm (item #19.2). Given that this multi-mental state measure also included tasks not assessed in the Wellman and Liu (2004) scale, we also examined scaling of items relating to emotion understanding, beliefs based on experiences, moral reasoning, interpretive ToM, and visual perspective taking. This broad battery allowed us to examine how these additional mental-state constructs scaled in comparison to traditional desires, diverse beliefs, false-beliefs, and sarcasm tasks assessed in assessed in Wellman and Liu (2004) and its derivatives (Bowman et al., 2017; Hiller et al., 2014; Peterson et al., 2012).

As seen in Figure 2, the difficulty estimates of items relating to emotion understanding (items #11.2 and #14.1), visual perspective taking (items #4.1 and #12.1), and moral reasoning (items #16.5 and #17.1) showed that the aspects of these constructs captured in our items developed in and among children's abilities to understand desires, diverse beliefs, false beliefs, and sarcasm. Understanding that someone would have negative emotions when presented with a previously aversive stimulus was the second easiest item for children to pass, after assessing understanding of common desires. Understanding a congruency between a person's stated positive emotion and their true inside emotion was more difficult, and fell between difficulty

levels for understanding diverse desires and diverse beliefs. Items assessing visual perspective taking also showed both relatively low difficulty (#4.1 was less difficult than diverse desires) and high difficulty (#12.1 was the most difficult for children to pass). Moral reasoning items (#16.5 and #17.1) both developed early, after common desires and before diverse desires. Interpretive ToM, the ability to take another's perspective in identifying an ambiguous image, developed alongside false belief understanding. Interestingly, questions about visual perspective taking (items #4.1 and #12.1) did not show a consistent rank-order in relation to other mental states, with #4.1 showing relatively a low difficulty and #12.1 showing the highest difficulty.

In addition to examining scaling of items, we also examined the range of difficulty scores for items across the measure more generally. Items were generally easy for participants to pass, in which 15 out of the 17 items had negative difficulty estimates. Item means were high, similarly illustrating a high proportion of participants passed each item, see Table 2. This constrained range in item performance may be problematic for extracting variability across the sample (e.g., if most participants are passing most measures, then there is less variability in ToM to correlate with other behavioral variables of interest). Additionally, high performance on this measure across the sample suggests that this measure may be better suited for a younger age range than studied (see further analyses and discussion below). However, low difficulty estimates are not problematic for evaluating discrimination and reliability. Although low difficulty estimates may lead researchers to think that a *measure* is overall not reliable, discrimination can evaluate how well individual items assess children's underlying ToM, and can be evaluated independently of the item's difficulty. In addition, item difficulty offers important information in creating the TIF in order to pinpoint the correct age range that is best assessed by this multi-mental state measure. Thus, low item difficulty does not prohibit further

exploration of the selected ToM measure, and measures may have items with low difficulty and still be reliable. An important benefit of the 2PL model, as opposed to the commonly employed Rasch and Guttman scaling methods, is that the 2PL model allows researchers to examine reliability (as indexed by the discrimination parameter) separately from item difficulty, which may lead researchers to believe that a measure is unreliable on the whole.

Assessing Item Discrimination

We examined item discrimination (*a*) extracted from each item's item-characteristic curve, and evaluated IRT parameters such that a 'good' measure should have items with high discrimination. High discrimination is indicated by a steeper slope, showing that the item is related to or predictive of the child's ToM, see Figure 2. Discriminating items can distinguish participants with similar but still differing ToM abilities; whereas low discriminating items (with values close to 0) suggest that an item is a poor assessment of ToM ability because participants with both low and high ToM ability would be equally likely to pass them. Discrimination estimates often range between 0.5 and 2.5 (Reeves & Fayers, 2005), although there is no definitive cut-off for evaluating "good" discrimination.

Analysis of the ToM Booklet 2 showed that most items had high discrimination, suggesting that items were able to differentiate or discriminate between participants with different ToM abilities. However, the true belief item (item #15.2) had a negative discrimination parameter, in which participants with less advanced ToM were likely to pass this item, and participants with more advanced ToM were more likely to fail this item. Therefore, the true belief item was unable to discriminate between different ToM abilities. Apart from this item, item discrimination estimates were good and ranged from 0.70 to 3.38.

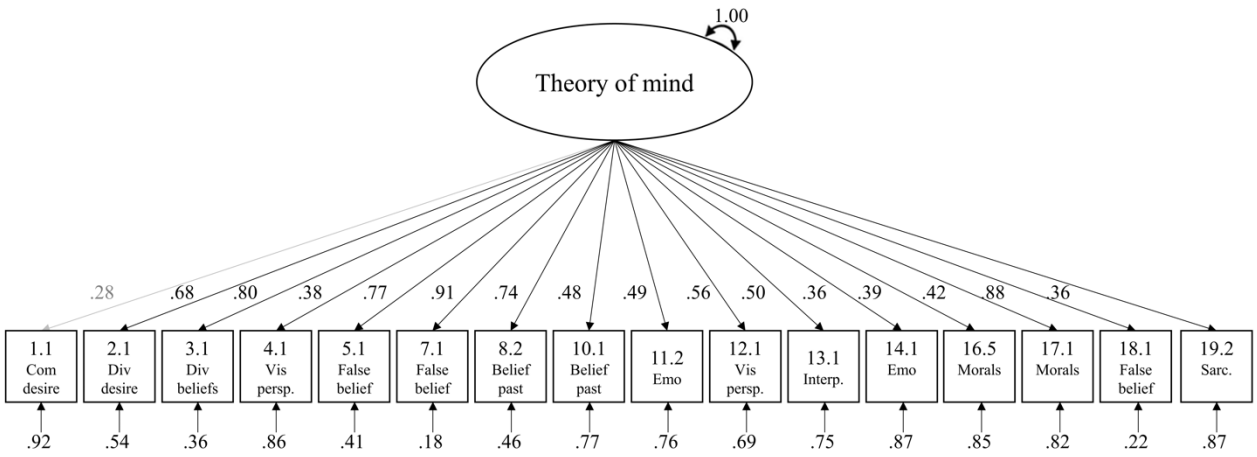
Using Difficulty and Discrimination to Revise the ToM Booklet 2

In examining item characteristic curves of the ToM Booklet 2, the item assessing true belief understanding showed high difficulty and poor discrimination, suggesting this item was a poor measure of children's ToM. We removed this true belief item (#15.2) from the larger measure, and evaluated the remaining 16 items.

Using the new subscale of 16 items, we evaluated the underlying latent factor structure of ToM. Following our analysis plan for confirming unidimensionality in 'Meeting IRT Model Assumptions,' we again evaluated acceptable fit criteria as a χ^2:df ratio ≤ 2 (Schreiber et al., 2006), Tucker-Lewis index (TLI) $\geq .90$ and comparative fit index (CFI) $\geq .90$ (Brown, 2006), and root mean square error of approximation (RMSEA) $\leq .08$ (Hu & Bentler, 1999). In this revised model including 16 items, a unidimensional model for ToM had good fit, in which the χ^2:df ratio = 1.30, TLI = .93, CFI = .94, and RMSEA = .04, 95% CI [0.02, 0.05], see Figure 3 for factor loadings and Table S2 for covariance matrix. The revised ToM Booklet 2 still met assumptions for both local independence (because we maintained the practice of selecting the first prediction item within each story), and unidimensionality demonstrated through good model fit.

Figure 3

Confirmatory Factor Analysis Depicting a Unidimensional ToM Latent Factor



Notes. Standardized loadings are shown for each indicator, and the variance of the latent factor was fixed to 1 in order to identify the model. Indicators corresponded to the following items: #1.1 – Common desires, #2.1 – Diverse desires, #3.1 – Diverse beliefs, #4.1 – Visual perspective taking, #5.1 – False belief, #7.1 – False belief, #8.2 – Belief based on past experience, #10.1 – Belief based on past experience, #11.2 – Emotion, #12.1 – Visual perspective taking, #13.1 – Interpretative ToM, #14.1 Emotion, #16.5 – Morals, #17.1 – Morals, #18.1 – False belief, #19.2 Sarcasm. Bolded factor loadings were significant, $p < .05$.

Assessing Reliability

We next examined the revised ToM storybook measure's reliability using two indices: Cronbach's α (as is common in the field) and the TIF extracted from the 2PL IRT model. These two indices differ in how they assess reliability: Cronbach's α describes how inter-related items are (i.e., how similar items across the scale are to each other), whereas the TIF describes how well the measure assesses ToM across different ToM abilities.

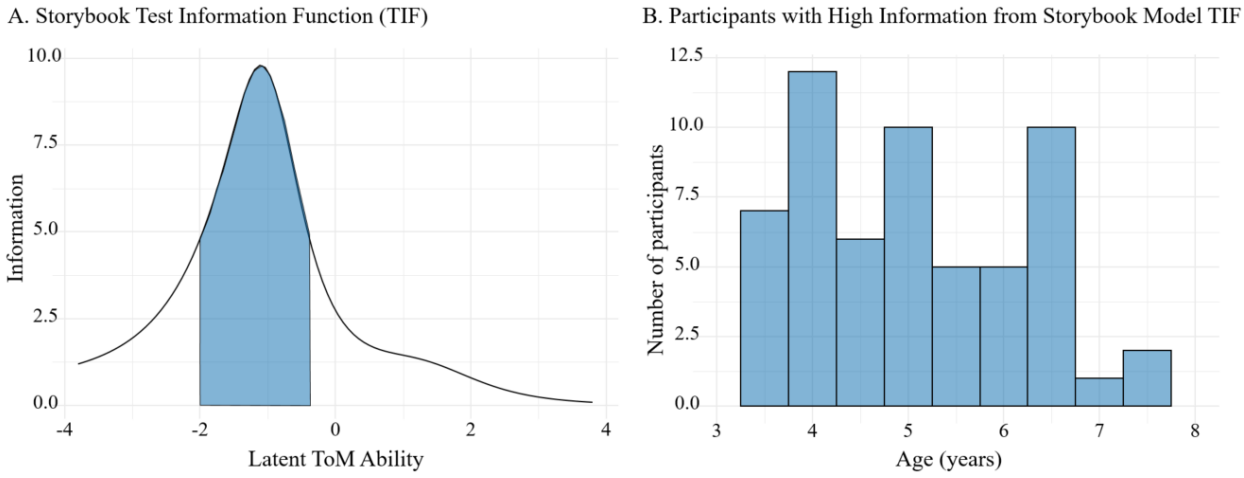
Internal consistency was in-line with existing ToM measures (see Beaudoin et al., 2020), $\alpha = 0.68$, suggesting that items were related. Note that we also examined α within the full set of

17 items, and scale modification suggestions for removing items: the 17-item composite returned a value of α = 0.64. Given the similarity in α-values for the 17- and 16-item composites, it would be difficult to justify removal of the true-belief item based on Cronbach's α alone. However, the discrimination parameter within the 2PL model gave clear evidence for removing this item from the composite.

Next, we examined the TIF for the revised 16-item ToM Booklet. The TIF illustrated that this measure best assessed children with low ToM ability. Specifically, the 16-item ToM Booklet had the highest information coverage for participants with a latent ToM ability (θ) between -0.5 and -2, see Figure 3A. To identify for whom this measure was best suited and to guide future research, we examined characteristics of participants with $-0.5 \leq \theta \geq -2$. Participants in this range (n = 58) were younger than the full sample (M_{age} = 5.05 years, SD = 1.12, range = 3.5 to 7.5 years), see Figure 3B.

Figure 3

Test Information Function (TIF) from the 2PL IRT Model and Age of Participants



Notes. Panel A: TIF for the 2 PL model, illustrating that this measure had the best coverage for children with a latent ToM score between -0.5 and -2 (shaded blue). Panel B: Histogram of the age of participants with a latent ToM ability between -0.5 and -2, in which the storybook had the highest information coverage.

Examining Items within the 16-item ToM Booklet with the Test Information Function (TIF)

As stated previously, selecting the correct ToM measure to capture variability in age ranges (particularly those outside of 3- to 5-years) is challenging. One aim of the present study was to determine the ToM abilities and participant ages for which the ToM Booklet 2 was best suited to assess. Current methods (e.g., examining the range and standard deviation of scores) do not allow researchers to pinpoint the ages for which the measure best captures variability. Our analysis of the TIF revealed that the 16-item ToM Booklet 2 was actually best able to assess ToM in children ages 3.5- to 7.5 years (not the full sample of 3.5- to 13-years-old).

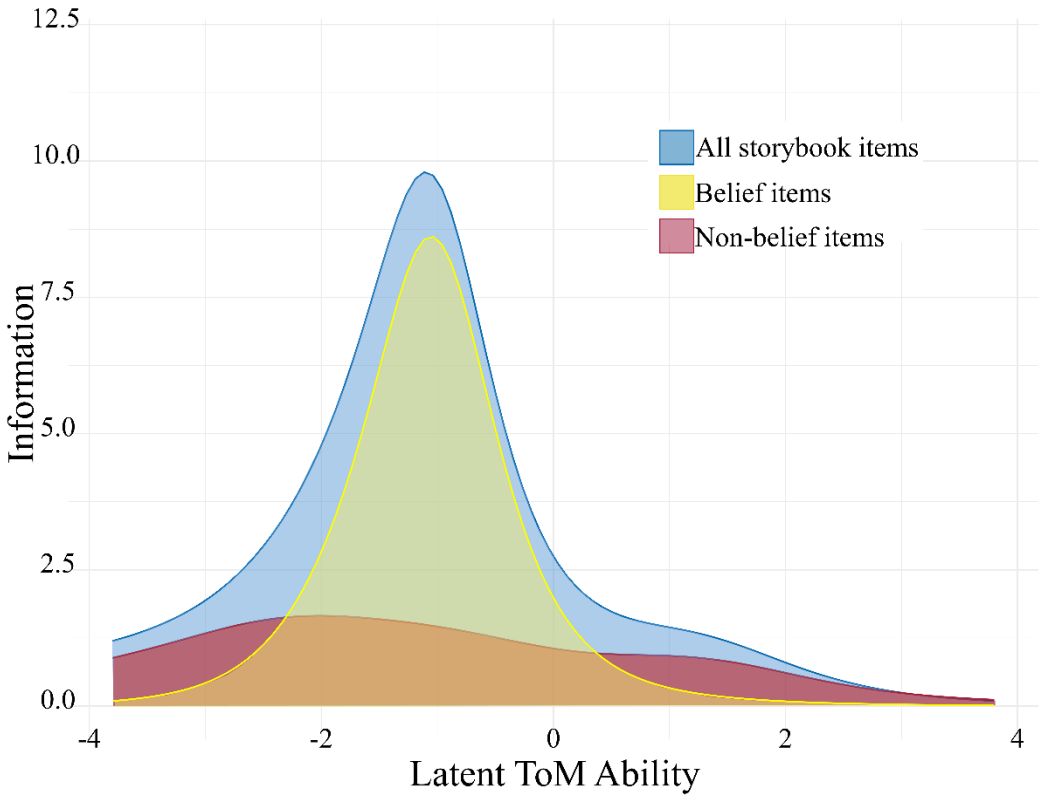
Next, we examined whether specific tasks were weighted more heavily in the measure's assessment of ToM given that this ToM measure included a broad diversity of mental states, but not in equal proportions: Whereas some mental states were reflected in a single item (e.g., understanding that people can have different desires), other mental states were reflected in multiple items (e.g., there were 3 items that assessed false-belief reasoning). Belief reasoning is highly studied within the field of ToM, as apparent in many reviews and meta-analyses (e.g., Devine & Hughes, 2014; Liu et al., 2008; Milligan et al., 2007; Wellman et al., 2001) and is often considered the hallmark of a mature and representational ToM (Astington & Baird, 2005; Leslie, 2000). Therefore, we examined whether a composite of only belief reasoning would be as informative as the total measure. Specifically, we fit a separate 2PL model to a composite of 6 belief items (Diverse beliefs: item #3.1; False belief: items #5.1, #7.1, #18.1; Belief based on

past experience: items #8.2, #10.1) to examine whether a composite of only belief items captured similar information compared to the full storybook measure. In addition, we fit a third 2PL model to a composite of only the 10 remaining non-belief items, to examine whether these other items increased the range of information provided by the full measure.

The TIF for the composite of 6 belief items (Figure 4, yellow curve) showed substantial overlap with the TIF for the full storybook measure (Figure 4, blue curve). This overlap supports the importance of including belief reasoning in larger composites of mental state reasoning. A scale of only non-belief items (Figure 4, red curve) showed a two-peaked distribution, in which some items were informative in the same range as belief understanding, and other items were more challenging, as shown by the lower apex being shifted toward higher latent ToM ability. Overall, the reduced peak for the non-belief items relative to the belief items reveals the overall comparatively greater amount of information (on children's ToM) provided by the belief items.

Figure 4

Test Information Function (TIF) from the Storybook and Belief Composite Measures



Notes. The TIFs from two 2PL models are shown for storybook items from the revised scale (in blue, 16 items), from a composite of only belief items (in yellow, 6 items), and from a composite of only non-belief items (in red, 10 items).

Assessing Validity

We assessed both relational validity (the relation with covariates) and construct validity (the relation with our target multi-mental state measure and other ToM measures) of the revised 16-item ToM Booklet 2. Specifically, participants also participated in measures of executive function and vocabulary, and therefore we examined relations between ToM and three common covariates that contribute to ToM development: age (Wellman & Liu, 2004), executive function (See meta-analysis: Devine & Hughes, 2014), and vocabulary (See meta-analysis: Milligan et al., 2007). In order to conduct correlations between variables, we created an average of each

participants' completed items out of the 16 items that were included in the final composite of items from the ToM Booklet 2. Because executive function and vocabulary measures were different across samples, we examined the Sacramento and Boston samples separately (see Table S3 for summary statistics for each sample).

Relational Validity. In both samples, ToM was correlated with age, (Sacramento: $r(95) = 0.46$, $p < .001$; Boston: $r(125) = 0.74$, $p < .001$), and vocabulary (Sacramento: $r(90) = 0.51$, $p < .001$; Boston: $r(91) = 0.50$, $p < .001$), in line with extant research, See Tables S4 and S5. However, in the Sacramento sample, ToM was not correlated with a conflict executive function measure (Up/Down, Kramer et al., 2015), nor with an emotional conflict executive function measure (Happy/Sad, Lagattuta et al., 2011), p's > .492, see Table S4. In a subset of 31 participants in the Boston sample, ToM was correlated with the pre- and post-switch trials of the DCCS ($r(29) = 0.39$, $p = .028$, see Table S5.

Construct Validity. We investigated whether our multi-mental state ToM measure correlated with two established questionnaire measures of ToM in the Sacramento sample. Specifically, we examined whether ToM was correlated with the TOMI-2 (Hutchins et al., 2014) and the CSUS (Tahiroglu et al., 2014), which are two measures of parents' reports of children's ToM which have been normed with this age group and have been shown to correlate with behavioral measures of children's ToM. Note that both of these measures showed high internal reliability (CSUS $\alpha = 0.90$, TOMI $\alpha = 0.96$), but IRT analyses could not be conducted because of reduced sample sizes in each of these measures.

In examining correlations between behavioral and parent-report ToM, the Pragmatics subscale within the TOMI-2 was correlated with children's ToM, $r(86) = 0.22$, $p = .037$, and the Intention ($r(81) = 0.22$, $p = .046$) and Perception ($r(72) = 0.27$, $p = .019$) CSUS subscales were

correlated with children's ToM. Thus, parent-report measures of children's ToM were generally related to children's behavioral performance on the ToM Booklet 2.

Discussion

The present study builds on a decade of concerted efforts in developmental science to improve research tools and practices by highlighting an under-utilized model for evaluating psychometric properties of ToM measures. Psychometric evaluations of ToM measures are particularly timely given the variability in ToM measures that exist, the breadth of concepts these measures aim to assess, and recent empirical evidence suggesting low reliability in existing ToM measures (Navarro et al., 2021; Warnell & Redcay, 2019). In this paper, we provide suggestions for undertaking such psychometric evaluations, and illustrate the use of Item Response Theory (IRT) to evaluate an existing multi-mental-state measure of ToM reasoning developed for a broad age range (preschool through middle childhood) (Richardson et al., 2018). Findings of our evaluation replicate and extend prior research examining how multiple mental state constructs (e.g., desires, beliefs) 'scale' across development (e.g., Wellman & Liu, 2004), inform future use of this specific ToM measure, and more generally outline an approach for evaluating existing and newly developed ToM measures in the field.

Insights from IRT Psychometric Evaluation of the ToM Booklet 2

We used a two-parameter logistic (2PL) IRT model to analyze the ToM Booklet 2 (Richardson et al., 2018)—a multi-item ToM measure that assesses children's understanding of common desires, diverse desires, diverse beliefs, false beliefs, beliefs based on past experiences, true beliefs, emotions, visual perspective taking, morals, interpretive ToM, and sarcasm. The measure was designed to be used with a broad age range and was previously used in a sample of children 3- to 12-years-old (Richardson et al., 2018). Findings demonstrated that, broadly, the

measure has good psychometric properties. In general, the items in the measure discriminated among different ToM abilities (assessed as a latent factor) and captured a reasonable range of difficulty levels. The true-belief item was an exception to this general pattern, and its item characteristic curve revealed poor discrimination among ToM abilities and a negative difficulty estimate (meaning that it was in fact more difficult to pass for children who had overall more ToM ability). We therefore removed this item from the measure, and further analyses revealed that the revised 16-item measure showed good reliability, and both relational and construct validity (as assessed by relations between children ToM performance and measures of age, executive functioning, vocabulary, and parent-report ToM). Thus, at the broadest level, our psychometric analysis reveals that the ToM Booklet 2 (in particular our revised 16-item version) is a reliable and valid measure of children's ToM.

The specific use of the 2PL IRT model, which yields *item characteristic curves* (which assess item-level difficulty and discrimination), and a *test information function* (*TIF;* which assess overall test reliability and appropriateness of the measure for the tested sample) also offered particular insights into ToM development and the use of this measure, some of which would not have been revealed with other more common psychometric models. First, Cronbach's α was unable to determine that the true-belief item was an unreliable ToM task, whereas the item characteristic curve extracted from the 2PL model clearly supported removing this item. Our finding that true-belief reasoning did not perform like other mental-state items is aligned with research demonstrating that true-belief understanding is distinct from other mental states in that it shows a unique non-linear developmental trajectory (Fabricius et al., 2010; Hedger & Fabricius, 2011; Schidelko et al., 2021), and is often used as a control or contrast condition for false belief rather than an index of mental-state reasoning in and of itself (e.g., Hyde et al., 2015;

Southgate et al., 2010). Our results raise intriguing questions about the nature of true-belief understanding in preschool and middle childhood, and suggest that it may not be appropriate to combine true-belief reasoning with other mental-state tasks in the formation of ToM composites. Second, we were able to use the TIF to reveal the optimal age-range for the use of this measure. The TIF also revealed that the belief items (which constituted 6 of the 16-items evaluated) accounted for the majority of information provided by the measure, although the inclusion of the additional non-belief items also importantly contributed to the variability captured in our sample. In the current measure, alternative explanations are possible as to why belief-items captured the most information. For example, these belief items may be critical in capturing meaningful variability in a particular aspect of children's ToM that is developing over this time period (e.g., development of the understanding that mental states are representations of reality, person-specific, and can be false; Leslie, 2000; Wellman, 2014). Alternatively, the mere over-representation of belief items in the measure may have accounted for the greater proportion of information captured by these items. Future measures are needed which contain a balanced number of items within each mental-state category in order to distinguish between these possibilities. Third, we conducted a factor analysis as part of our test of assumptions for the appropriate use of IRT. This analysis revealed that a unidimensional model, in which all items loaded onto a single latent ToM factor, fit the data best. This is one of the few studies that have examined the factor structure of ToM with the traditional tasks used in preschool (e.g., Wellman & Liu, 2004), in addition to measures administered in middle-childhood (e.g., interpretive ToM, sarcasm), with items that are matched in task format. Thus, the present study supports the possibility that ToM is a single cognitive construct that can be measured through a variety of mental-state tasks.

Thus, our use of IRT to evaluate the psychometric properties of this ToM Booklet 2 offered insights that have theoretical implications for future research and ToM development. More directly, psychometric findings offer clear guidelines for future use of this measure: (1) the true belief item (#15.2) should be removed, (2) the measure is likely optimal for use with children 3.5 to 7.5 years of age, and (3) the scale appears to best capture ToM as reflected in performance on belief items, however the inclusion of additional non-belief items is important for assessing maximal variability. More balanced measures with equal representation of multiple mental states are needed for future research.

'Scaling' Multiple Mental-State Constructs over Preschool to Middle Childhood

Our psychometric analysis of the ToM Booklet 2 (Richardson et a., 2018) also offers new insight into a developmental hierarchy of different mental-state constructs, expanding prior research on the 'scaling' of ToM (e.g., Wellman & Liu, 2004; Osterhaus et al. 2016; 2022; Shahaeian et al., 2011). First, we replicated decades of research that have found that children can accurately reason about desires before they can accurately reason about diverse beliefs, which precedes accurate false-belief reasoning and understanding of sarcasm (Happé, 1994; Wellman & Liu, 2004). This conceptual replication is particularly important in order to validate this recently-developed measure. Critically, we also provide novel findings that demonstrate how additional mental-state constructs fit within this more established developmental hierarchy. We note that there is a need for additional research replicating our results with novel items, and our analyses here are limited by the specific items used in the measure to represent a given mental-state construct, which were mostly represented by one or two items. Nonetheless, our results offer intriguing insights that extend prior research on a hierarchy of mental-state development.

Specifically, reasoning about emotions developed after common desires and before diverse beliefs. The negative emotion item was easier for children to pass than the positive emotion item, in line with research showing that children and parents discuss negative emotions and causes of emotions more when discussing *negative* versus positive emotions (Lagattuta & Wellman, 2002). Moral reasoning similarly developed amidst desire understanding, which was expected given that items assessed basic rule-following principles, such as that stealing is naughty, which tend to develop earlier in preschool on a similar timeline to development of desire-understanding (e.g., by 2.5-years-old, Dahl & Killen, 2018; Smetana & Braeges, 1990; Smetana et al., 2012). Interpretive ToM developed alongside false belief, which was consistent with research showing that interpretive ToM is a later-developing mental state (Lagattuta et al., 2010; Carpendale & Chandler, 1996). Although some studies have shown that interpretive ToM develops in early middle-childhood (about 6- to 7-years-old, e.g., Carpendale & Chandler, 1996; Lagattuta et al., 2010), we could not determine whether our effect of interpretive ToM developing alongside false belief was robust given that there was only one interpretive ToM item administered. Further assessments of both interpretation and false belief are necessary in order to determine whether these mental states develop concurrently, or whether interpretive ToM may develop later than ToM depending on task demands.

There were other mental-state constructs tested in which items were both easier and more difficult for children to pass. Specifically, visual perspective taking developed both before diverse desires, as well as between diverse beliefs and false belief. This rank-order is consistent with research showing perspective-taking development in early childhood (Carlson et al., 2004; Moll & Meltzoff, 2011), but also suggests that there are nuances in visual perspective taking tasks that influences task difficulty and may be related to developments in underlying mental-

state reasoning or may be related to developments in constructs that support perspective-taking (e.g., executive function). For example, visual perspective taking is thought of as a precursor to understanding others' knowledge states because it affords opportunities to learn that knowledge states are dependent upon what one is able to see (Bigelow & Dugas, 2008). Although the ToM Booklet 2 did not include a knowledge item, one of our perspective-taking tasks fits in this scale where knowledge would likely be based on prior research (between diverse beliefs and false belief, Wellman & Liu, 2004). Finally, beliefs based on past experiences developed in and around false belief items, including before and after. These additional tasks show nuances in the developmental trajectory for belief reasoning beyond the field's canonical false-belief task, and are aligned with research showing that reasoning about past influences on mental states is more challenging than reasoning about concurrent influences; this type of belief-reasoning has a protracted trajectory, with even 6- and 7-year-olds struggling with reasoning about the influence of memory and past experiences (Lagattuta & Wellman, 2001; see review: Lagattuta, 2014).

Limitations to the Present Study

The present study represents a foundation for assessing novel social cognitive measures using the 2PL model within an item response theory framework. We acknowledge limitations of the present study and offer recommendations for future research. First, in order to meet assumptions for IRT analyses and in order to use the same scoring scheme across laboratories, we only evaluated prediction questions, in which children were asked to predict a character's action based on their mental state; we did not evaluate explanation questions within this measure. We chose to select prediction questions because explanations questions are not consistently included in ToM batteries, and prediction scores are binary and are not affected by differences across coding schemes across laboratories. Future research may consider other models, such as

testlet IRT models, to account for the non-independence of items, or consider designing ToM tasks that already meet assumptions of local independence.

Second, as noted above, we only had one representation of several mental states – specifically, common desires, diverse desires, diverse beliefs, interpretive ToM, and sarcasm. Therefore, we cannot draw conclusions beyond the specific aspects of the construct that each particular item tests, and we encourage future research to include additional items testing a given construct to further examine both nuances in the developmental trajectory of mental states as well as to more robustly test the construct's scaled position over multiple items. Specifically, there are open questions about whether our results illustrating that a subset of only belief items captured the majority of information about children's ToM (as illustrated in the TIF) is because of the importance of belief-understanding in ToM development (e.g., in that belief understanding assesses person-specific and reality-independent aspects of mental state reasoning which undergo much development over early to middle childhood), or whether the disproportional amount of information is due to the overrepresentation of belief items in the composite more generally.

Finally, although our sample was diverse in age, geographic region, and race and ethnicity, our sample size was close to the minimum required for latent factor analyses (Boomsma & Hoogland, 2001; Hoogland & Boomsma, 1998; Kline, 2015), and ideally IRT analyses should include samples large enough to reflect the population parameters. Thus, further replication is needed with larger samples.

Conclusion

In sum, there is a need for validated and reliable ToM measures for children, particularly to assess ToM beyond the preschool years. We illustrate a process for validating newly-developed scales and apply this method to a recently-published measure of ToM developed for a

wide age range. Our findings inform steps forward in producing reliable measures and replicable research in developmental science generally, and also inform how to use this measure in future research, such as excluding true belief from larger composites of ToM and using the selected measure in a specific age range (3.5- to 7.5-years-old). Results revealed that reliable composites can be developed and used that include diversity in mental state reasoning, spanning discrepant mental state processes such as desires, beliefs, moral reasoning, emotions, and sarcasm. Importantly, we highlight the utility of the 2PL IRT model in designing and evaluating robust and reliable measures of children's ToM, and are the first study to utilize the 'discrimination' parameter within an IRT model in evaluating a ToM measure. This analysis demonstrated the utility of the ToM booklet 2 (Richardson et al., 2018) as a reliable measure of children's ToM with good discriminability, and thus the field can have confidence in the use of this in future individual differences studies of ToM.

References

Altschuler, M., Sideridis, G., Kala, S., Warshawsky, M., Gilbert, R., Carroll, D., Burger-Caplan, R., & Faja, S. (2018). Measuring individual differences in cognitive, affective, and spontaneous theory of mind among school-aged children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 48,* 3945–3957. https://doi.org/10.1007/s10803-018-3663-1

Astington, J. W., & Baird, J. A. (Eds.). (2005). *Why language matters for theory of mind*. Oxford University Press.

Bambha, V. P., & Casasola, M. (2021). From lab to zoom: Adapting training study methodologies to remote conditions. *Frontiers in Psychology, 12,* 694728. https://doi.org/10.3389/fpsyg.2021.694728

Beaudoin, C., Leblanc, E., Gagner, C., & Beauchamp, M. H. (2020). Systematic review and inventory of theory of mind measures for young children. *Frontiers in Psychology, 10,* 2905. https://doi.org/10.3389/fpsyg.2019.02905

Bigelow, A. E., & Dugas, K. (2008). Relations among preschool children's understanding of visual perspective taking, false belief, and lying. *Journal of Cognition and Development, 9*(4), 411-433. https://doi.org/10.1080/15248370802678299

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition, 77*(1), B25-B31. https://doi.org/10.1016/S0010-0277(00)00096-2

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. *Structural equation models: Present and future. A Festschrift in honor of Karl Jöreskog, 2*(3), 139-168.

Bowman, L. C., & Wellman, H. M. (2014). Neuroscience contributions to childhood theory of mind development. In Saracho, O. N. (Ed.), *Contemporary perspectives on research in theory of mind in early childhood education* (pp. 195-223). Information Age Publishing.

Bowman, L. C., Kovelman, I., Hu, X., & Wellman, H. M. (2015). Children's belief-and desire-reasoning in the temporoparietal junction: evidence for specialization from functional near-infrared spectroscopy. *Frontiers in Human Neuroscience, 9*, 560. https://doi.org/10.3389/fnhum.2015.00560

Bowman, L. C., Liu, D., Meltzoff, A. N., & Wellman, H. M. (2012). Neural correlates of belief- and desire-reasoning in 7-and 8-year-old children: An event-related potential study. *Developmental Science, 15*(5), 618-632. https://doi.org/10.1111/j.1467-7687.2012.01158.x

Bowman, L. C., Thorpe, S. G., Cannon, E. N., & Fox, N. A. (2017). Action mechanisms for social cognition: Behavioral and neural correlates of developing theory of mind. *Developmental Science, 20*(5), e12447. https://doi.org/10.1111/desc.12447

Carlson, S. M., Claxton, L. J., & Moses, L. J. (2015). The relation between executive function and theory of mind is more than skin deep. *Journal of Cognition and Development, 16*(1), 186-197. https://doi.org/10.1080/15248372.2013.824883

Carlson, S. M., Koenig, M. A., & Harms, M. B. (2013). Theory of mind. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(4), 391-402. https://doi.org/10.1002/wcs.1232

Carlson, S. M., Mandell, D. J., & Williams, L. (2004). Executive function and theory of mind: Stability and prediction from ages 2 to 3. *Developmental Psychology, 40*(6), 1105–1122. https://doi.org/10.1037/0012-1649.40.6.1105

Carlson, S. M., Mandell, D. J., & Williams, L. (2004). Executive function and theory of mind: stability and prediction from ages 2 to 3. *Developmental Psychology, 40*(6), 1105–1122. https://doi.org/10.1037/0012-1649.40.6.1105

Carpendale, J. I., & Chandler, M. J. (1996). On the distinction between false belief understanding and subscribing to an interpretive theory of mind. *Child Development, 67*(4), 1686-1706. https://doi.org/10.1111/j.1467-8624.1996.tb01821.x

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Choi, Y. J., and Luo, Y. (2015). 13-month-olds' understanding of social interactions. *Psychological Science, 26*, 274–283. https://doi.org/10.1177/0956797614562452

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science, 1*(3), 140216. https://doi.org/10.1098/rsos.140216

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science, 4*(12), 171085. https://doi.org/10.1098/rsos.171085

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Dahl, A., & Killen, M. (2018). Moral reasoning: Theory and research in developmental science. In J. Wixted (Ed.), *The Steven's Handbook of Experimental Psychology and Cognitive Neuroscience, Vol. 3: Developmental and Social Psychology* (S. Ghetti, Vol. Ed.)., 4th edition. New York: Wiley.

Devine, R. T., & Apperly, I. A. (2021). Willing and able? Theory of mind, social motivation, and social competence in middle childhood and early adolescence. *Developmental Science, 25*(1), e13137. https://doi.org/10.1111/desc.13137

Devine, R. T., & Hughes, C. (2013). Silent films and strange stories: Theory of mind, gender, and social experiences in middle childhood. *Child Development, 84*(3), 989-1003. https://doi.org/10.1111/cdev.12017

Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child Development, 85*(5), 1777–1794. https://doi.org/10.1111/cdev.12237

Devine, R. T., & Hughes, C. (2016). Measuring theory of mind across middle childhood: Reliability and validity of the silent films and strange stories tasks. *Journal of Experimental Child Psychology, 149*, 23-40. https://doi.org/10.1016/j.jecp.2015.07.011

Devine, R. T., White, N., Ensor, R., & Hughes, C. (2016). Theory of mind in middle childhood: Longitudinal associations with executive function and social competence. *Developmental Psychology, 52*(5), 758. https://doi.org/10.1037/dev0000105

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology, 105(3), 399-412. https://doi.org/10.1111/bjop.12046

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody picture vocabulary test*. Pearson Assessments.

Ebert, S., Peterson, C., Slaughter, V., & Weinert, S. (2017). Links among parents' mental state language, family socioeconomic status, and preschoolers' theory of mind development. *Cognitive Development, 44,* 32-48. https://doi.org/10.1016/j.cogdev.2017.08.005

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fabricius, W. V., Boyer, T. W., Weimer, A. A., & Carroll, K. (2010). True or false: Do 5-year-olds understand belief? *Developmental Psychology, 46*(6), 1402–1416. https://doi.org/10.1037/a0017648

Fink, E., de Rosnay, M., Patalay, P., & Hunt, C. (2020). Early pathways to bullying: A prospective longitudinal study examining the influences of theory of mind and social preference on bullying behavior during the first 3 years of school. *British Journal of Developmental Psychology, 38*(3), 458-477. https://doi.org/10.1111/bjdp.12328

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science, 3*(4), 484-501. https://doi.org/10.1177/2515245920951747

Grosse Wiesmann, C., Schreiber, J., Singer, T., Steinbeis, N., & Friederici, A. D. (2017). White matter maturation is associated with the emergence of theory of mind in early childhood. *Nature Communications, 8,* 14692. https://doi.org/10.1038/ncomms14692

Guajardo, N. R., Petersen, R., and Marshall, T. R. (2013). The roles of explanation and feedback in false belief understanding: A microgenetic analysis. *Journal of Genetic Psychology, 174*, 225–252. https://doi.org/10.1080/00221325.2012.682101

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23-34. https://doi.org/10.20982/tqmp.08.1.p023

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods, 12*(3), 179-185. https://doi.org/10.1038/nmeth.3288

Hanson, L. K., Atance, C. M., & Paluck, S. W. (2014). Is thinking about the future related to theory of mind and executive function? Not in preschoolers. *Journal of Experimental Child Psychology, 128*, 120-137. https://doi.org/10.1016/j.jecp.2014.07.006

Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders, 24*(2), 129-154. https://doi.org/10.1007/BF02172093

Hedger, J. A., & Fabricius, W. V. (2011). True belief belies false belief: Recent findings of competence in infants and limitations in 5-year-olds, and implications for theory of mind development. *Review of Philosophy and Psychology, 2*(3), 429-447. https://doi.org/10.1007/s13164-011-0069-9

Henning, A., Spinath, F. M., & Aschersleben, G. (2011). The link between preschoolers' executive function and theory of mind and the role of epistemic states. *Journal of*

Experimental Child Psychology, *108*(3), 513-531. https://doi.org/10.1016/j.jecp.2010.10.006

Henry, J. D., Phillips, L. H., Ruffman, T., & Bailey, P. E. (2013). A meta-analytic review of age differences in theory of mind. *Psychology and Aging*, *28*(3), 826–839. https://doi.org/10.1037/a0030677

Hiller, R. M., Weber, N., & Young, R. L. (2014). The validity and scalability of the theory of mind scale with toddlers and preschoolers. *Psychological Assessment, 26*(4), 1388. https://doi.org/10.1037/a0038320

Holl, A. K., Kirsch, F., Rohlf, H., Krahé, B., & Elsner, B. (2018). Longitudinal reciprocity between theory of mind and aggression in middle childhood. *International Journal of Behavioral Development, 42*(2), 257-266. https://doi.org/10.1177/0165025417727875

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329-367. https://doi.org/10.1177%2F0049124198026003003

Hori, K., Fukuhara, H., & Yamada, T. (2020). Item response theory and its applications in educational measurement part I: Item response theory and its implementation in R. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(2), e1531. https://doi.org/10.1002/wics.1531

Hutchins, T. L., Prelock, P. A., & Bouyea, L. B. (2014). Technical manual for the theory of mind inventory-2. http://www.theoryofmindinventory.com/task-battery.

Hwa-Froelich, D. A., Matsuo, H., & Jacobs, K. (2017). False belief performance of children adopted internationally. *American Journal of Speech-Language Pathology, 26*(1), 29-43. https://doi.org/10.1044/2016_AJSLP-15-0152

Hyde, D. C., Aparicio Betancourt, M., & Simon, C. E. (2015). Human temporal-parietal junction spontaneously tracks others' beliefs: A functional near-infrared spectroscopy study. *Human Brain Mapping*, *36*(12), 4831-4846. https://doi.org/10.1002/hbm.22953

Iannotti, R. J. (1978). Effect of role-taking experiences on role taking, empathy, altruism, and aggression. *Developmental Psychology, 14*, 119 –124. http://dx.doi.org/10.1037/0012-1649.14.2.119

Imuta, K., Henry, J. D., Slaughter, V., Selcuk, B., & Ruffman, T. (2016). Theory of mind and prosocial behavior in childhood: A meta-analytic review. *Developmental Psychology, 5*2(8), 1192–1205. https://doi.org/10.1037/dev0000140

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences, 29,* 105-110. https://doi.org/10.1016/j.cobeha.2019.04.010

Jin, X., Li, P., He, J., & Shen, M. (2017). Cooperation, but not competition, improves 4-year-old children's reasoning about others' diverse desires. *Journal of Experimental Child Psychology, 157,* 81-94. https://doi.org/10.1016/j.jecp.2016.12.010

Kaufman, A. S. (2004). Kaufman brief intelligence test–second edition (KBIT-2). *Circle Pines, MN: American Guidance Service*.

Killen, M., & Smetana, J.G. (2013). *Handbook of moral development*. Psychology Press. https://doi.org/10.4324/9780203581957

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Kramer, H. J., Lagattuta, K. H., & Sayfan, L. (2015). Why is happy–sad more difficult? Focal emotional information impairs inhibitory control in children and adults. *Emotion, 15*(1), 61–72. https://doi.org/10.1037/emo0000023

Kuhnert, R. L., Begeer, S., Fink, E., & de Rosnay, M. (2017). Gender-differentiated effects of theory of mind, emotion understanding, and social preference on prosocial behavior development: A longitudinal study. *Journal of Experimental Child Psychology, 154*, 13–27. https://doi.org/10.1016/j.jecp.2016.10.001

Lagattuta, K. H. (2014). Linking past, present, and future: Children's ability to connect mental states and emotions across time. *Child Development Perspectives*, *8*(2), 90-95. https://doi.org/10.1111/cdep.12065

Lagattuta, K. H., & Wellman, H. M. (2001). Thinking about the past: Early knowledge about links between prior experience, thinking, and emotion. *Child Development*, *72*(1), 82-102. https://doi.org/10.1111/1467-8624.00267

Lagattuta, K. H., & Wellman, H. M. (2002). Differences in early parent-child conversations about negative versus positive emotions: implications for the development of psychological understanding. *Developmental Psychology*, *38*(4), 564–580. https://doi.org/10.1037/0012-1649.38.4.564

Lagattuta, K. H., Sayfan, L., & Blattman, A. J. (2010). Forgetting common ground: six-to seven-year-olds have an overinterpretive theory of mind. *Developmental Psychology*, *46*(6), 1417. https://doi.org/10.1016/bs.acdb.2014.11.005

Lagattuta, K. H., Sayfan, L., & Monsour, M. (2011). A new measure for assessing executive function across a wide age range: children and adults find happy-sad more difficult than day-night. *Developmental Science*, *14*(3), 481-489. https://doi.org/10.1111/j.1467-7687.2010.00994.x

Lecce, S., Bianco, F., Devine, R. T., Hughes, C. (2017). Relations between theory of mind and executive function in middle childhood: A short-term longitudinal study. *Journal of Experimental Child Psychology, 163*, 69-86. https://doi.org/10.1016/j.jecp.2017.06.011

Leslie, A. M. (2000). How to acquire a 'representational theory of mind'. Metarepresentations: A multidisciplinary perspective, 197-223.

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of mind development in Chinese children: A meta-analysis of false-belief understanding across cultures and languages. Developmental Psychology, 44(2), 523–531. https://doi.org/10.1037/0012-1649.44.2.523

Lüdecke D (2023). sjPlot: Data Visualization for Statistics in Social Science. R package version 2.8.14, https://CRAN.R-project.org/package=sjPlot

McHugh M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

McNeish D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23(*3), 412–433. https://doi.org/10.1037/met0000144

Metsämuuronen, J. (2022). Attenuation-corrected estimators of reliability. *Applied Psychological Measurement, 46*(8), 720-737. https://doi.org/10.1177/01466216221108131

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development, 78*(2), 622–646. https://doi.org/10.1111/j.1467-8624.2007.01018.x

Min, S., Aryadoust, V. (2021). A systematic review of item response theory in language assessment: Implications for the dimensionality of language ability. *Studies in Educational Evaluation, 68*, 100963. https://doi.org/10.1016/j.stueduc.2020.100963

Moll, H., & Meltzoff, A. N. (2011). How does it look? Level 2 perspective-taking at 36 months of age. *Child Development, 82*(2), 661-673. https://doi.org/10.1111/j.1467-8624.2010.01571.x

Mull, M. S., & Evans, E. M. (2010). Did she mean to do it? Acquiring a folk theory of intentionality. *Journal of Experimental Child Psychology*, *107*(3), 207-228. https://doi.org/10.1016/j.jecp.2010.04.001

Müller, U., Liebermann-Finestone, D. P., Carpendale, J. I., Hammond, S. I., & Bibok, M. B. (2012). Knowing minds, controlling actions: The developmental relations between theory of mind and executive function from 2 to 4 years of age. *Journal of Experimental Child Psychology, 111*(2), 331-348. https://doi.org/10.1016/j.jecp.2011.08.014

Muris, P., Steerneman, P., Meesters, C., Merckelback, H., Horselenberg, R., van den Hogen, T., & van Dongen, L. (1999). The TOM test: A new instrument for assessing theory of mind in mormal children and children with pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 29*, 67–80. https://doi.org/10.1023/A:1025922717020

Nilsson, K. K., & de Lopez, K. M. J. (2016). Theory of mind in children with specific language impairment: A systematic review and meta-analysis. *Child Development*, *87*(1), 143-153. https://doi.org/10.1111/cdev.12462

Nimon, K., Zientek, L. R., & Henson, R. K. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Psychology*, *3*, 102. https://doi.org/10.3389/fpsyg.2012.00102

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M.,

Scherer, L. D., Schönbrodt, F.D., Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73, 719-748. https://doi.org/10.1146/annurev-psych-020821-114157

Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill

O'Kearney, R., Salmon, K., Liwag, M., Fortune, C. A., & Dawel, A. (2017). Emotional abilities in children with oppositional defiant disorder (ODD): Impairments in perspective-taking and understanding mixed emotions are associated with high callous–unemotional traits. *Child Psychiatry & Human Development*, *48*, 346-357. https://doi.org/10.1007/s10578-016-0645-4

Olineck, K. M., & Poulin-Dubois, D. (2007). Imitation of intentional actions and internal state language in infancy predict preschool theory of mind skills. *European Journal of Developmental Psychology, 4*(1), 14-30. https://doi.org/10.1080/17405620601046931

Ong, D. C., Zaki, J., & Goodman, N. D. (2019). Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in Cognitive Science, 11*(2), 338-357. https://doi.org/10.1111/tops.12371

Osterhaus, C., Koerber, S. and Sodian, B. (2016), Scaling of advanced theory-of-mind tasks. *Child Development, 87*(6), 1971-1991. https://doi.org/10.1111/cdev.12566

Osterhaus, C., Kristen-Antonow, S., Kloo, D., & Sodian, B. (2022). Advanced scaling and modeling of children's theory of mind competencies: Longitudinal findings in 4- to 6-year-olds. International Journal of Behavioral Development, 46(3), 251-259. https://doi.org/10.1177/01650254221077334

Padilla, M. A., & Veprinsky, A. (2012). Correlation attenuation due to measurement error: A new approach using the bootstrap procedure. *Educational and Psychological Measurement, 72*(5), 827–846. https://doi.org/10.1177/0013164412443963

Peterson, C. C., & Wellman, H. M. (2009). From fancy to reason: Scaling deaf and hearing children's understanding of theory of mind and pretense. *British Journal of Developmental Psychology, 27*(2), 297–310. https://doi.org/10.1348/026151008X299728

Peterson, C. C., Wellman, H. M., & Liu, D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Development, 76*(2), 502-517. https://doi.org/10.1111/j.1467-8624.2005.00859.x

Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Development, 83*(2), 469-485. https://doi.org/10.1111/j.1467-8624.2011.01728.x

Phillips, A. T., and Wellman, H. M. (2005). Infants' understanding of object directed action. *Cognition, 98,* 137–155. https://doi.org/10.1016/j.cognition.2004.11.005

Pluck, G., Córdova, M. A., Bock, C., Chalen, I., & Trueba, A. F. (2021). Socio-economic status, executive functions, and theory of mind ability in adolescents: Relationships with language ability and cortisol. *British Journal of Developmental Psychology, 39*(1), 19-38. https://doi.org/10.1111/bjdp.12354

Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science, 15*(2), 384-396. https://doi.org/10.1177/1745691619896607

R Core Team. (2019). R (Version 3.6.2.). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. In *Assessing Quality of Life in Clinical Trials: Methods of Practice, 2,* 55-73.

Revelle, W. (2019). psych: Procedures for psychological, psychometric, and personality research. Northwestern University, Evanston, Illinois. R package version 1.8.12, https://CRAN.R-project.org/package=psych

Richardson, H., & Saxe, R. (2020). Development of predictive responses in theory of mind brain regions. *Developmental Science, 23*(1), e12863. https://doi.org/10.1111/desc.12863

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications, 9*(1), 1027. https://doi.org/10.1038/s41467-018-03399-2

Rizopoulos, D. (2006). ltm: an R package for latent variable modeling and item response analysis. *Journal of Statistical Software*, *17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Savalei, V., & Reise, S. P. (2019). Don't forget the model in your model-based reliability coefficients: A reply to McNeish (2018). *Collabra: Psychology*, *5*(1). https://doi.org/10.1525/collabra.247

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science, 17*(8), 692-699. https://doi.org/10.1111/j.1467-9280.2006.01768.x

Schidelko, L. P., Schünemann, B., Rakoczy, H., & Proft, M. (2021). Online testing yields the same results as lab testing: A validation study with the false belief task. *Frontiers in Psychology, 12*, 703238. https://doi.org/10.3389/fpsyg.2021.703238

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews, 42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009

Shahaeian, A., Peterson, C. C., Slaughter, V., & Wellman, H. M. (2011). Culture and the sequence of steps in theory of mind development. *Developmental Psychology, 47*(5), 1239–1247. https://doi.org/10.1037/a0023899

Slaughter, V. (2015). Theory of mind in infants and young children: A review. *Australian Psychologist, 50*(3), 169-172. https://doi.org/10.1111/ap.12080

Slaughter, V., Imuta, K., Peterson, C. C., & Henry, J. D. (2015). Meta-analysis of theory of mind and peer popularity in the preschool and early school years. *Child Development, 86*(4), 1159-1174. https://doi.org/10.1111/cdev.12372

Smetana, J. G., & Braeges, J. L. (1990). The development of toddlers' moral and conventional judgments. *Merrill-Palmer Quarterly, 36*(3), 329-346.

Smetana, J. G., Jambon, M., Conry-Murray, C., & Sturge-Apple, M. L. (2012). Reciprocal associations between young children's developing moral judgments and theory of mind. *Developmental Psychology, 48*(4), 1144–1155. https://doi.org/10.1037/a0025891

Sodian, B., Kristen-Antonow, S., & Kloo, D. (2020). How does children's theory of mind become explicit? A review of longitudinal findings. *Child Development Perspectives, 14*(3), 171-177. https://doi.org/10.1111/cdep.12381

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, *13*(6), 907-912. https://doi.org/10.1111/j.1467-7687.2009.00946.x

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 73-191.

Stellwagen, K. K., & Kerig, P. K. (2013). Dark triad personality traits and theory of mind among school-age children. *Personality and Individual Differences, 54*(1), 123–127. https://doi.org/10.1016/j.paid.2012.08.019

Tahiroglu, D., Moses, L. J., Carlson, S. M., Mahy, C. E. V., Olofson, E. L., & Sabbagh, M. A. (2014). The Children's Social Understanding Scale: Construction and validation of a parent-report measure for assessing individual differences in children's theories of mind. *Developmental Psychology, 50*(11), 2485–2497. https://doi.org/10.1037/a0037914

Tahiroglu, D., Moses, L. J., Carlson, S. M., Mahy, C. E., Olofson, E. L., & Sabbagh, M. A. (2014). The children's social understanding scale: Construction and validation of a parent-report measure for assessing individual differences in children's theories of mind. *Developmental Psychology, 50*(11), 2485. https://doi.org/10.1037/a0037914

Taylor, M., Cartwright, B. S., & Bowden, T. (1991). Perspective taking and theory of mind: Do children predict interpretive diversity as a function of differences in observers' knowledge? *Child Development, 62*(6), 1334-1351. https://doi.org/10.1111/j.1467-8624.1991.tb01609.x

Wang, S., Andrews, G., Pendergast, D., Neumann, D., Chen, Y., & Shum, D. H. K. (2021). A cross-cultural study of theory of mind using strange stories in school-aged children from Australia and Mainland China. *Journal of Cognition and Development, 23*(1), 40-63. https://doi.org/10.1080/15248372.2021.1974445

Wang, Z., Devine, R. T., Wong, K. K., & Hughes, C. (2016). Theory of mind and executive function during middle childhood across cultures. *Journal of Experimental Child Psychology, 149*, 6-22. https://doi.org/10.1016/j.jecp.2015.09.028

Warnell, K. R., & Redcay, E. (2019). Minimal coherence among varied theory of mind measures in childhood and adulthood. *Cognition, 191*. https://doi.org/10.1016/j.cognition.2019.06.009

Wellman, H. M. (2014). *Making minds: How theory of mind develops.* Oxford University Press.

Wellman, H. M. (2018). Theory of mind across the lifespan?. *Zeitschrift für Psychologie*.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*(2), 523-541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655-684. https://doi.org/10.1111/1467-8624.00304

Wellman, H. M., Fang, F., & Peterson, C. C. (2011). Sequential progressions in a theory-of-mind scale: Longitudinal perspectives. *Child Development, 82*(3), 780-792. https://doi.org/10.1111/j.1467-8624.2011.01583.x

Wellman, H. M., Fang, F., Liu, D., Zhu, L., & Liu, G. (2006). Scaling of theory-of-mind understandings in Chinese children. *Psychological Science, 17*(12), 1075–1081. https://doi.org/10.1111/j.1467-9280.2006.01830.x

Wilson, J., Andrews, G., Hogan, C., Wang, Si., Shum, D. H. K. (2018). Executive function in middle childhood and the relationship with theory of mind. *Developmental Neuropsychology, 43*(3), 163-182. https://doi.org/10.1080/87565641.2018.1440296

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103-128. https://doi.org/10.1016/0010-0277(83)90004-5

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols, 1*(1), 297-301. https://doi.org/10.1038/nprot.2006.46

Appendices

Appendix S1. Equating In-Person and Remote Testing Experiences

The task protocol was modified slightly to ensure that children who participated in the study remotely via Zoom understood instructions and could properly view the stimuli. Stimuli were presented via screenshare on a Powerpoint slide by the experimenter. Before the testing procedure, the participant's caregiver set up their Zoom screen so that the stimuli were in full screen mode, the participant's video was hidden (so that children were not distracted by their own video), and the experimenter's video was to the right side of the screen so that stimuli were not obtruded by the experimenter. If the participant was comfortable wearing headphones, the experimenter requested that the participant wear them to prevent being distracted by other noises in the home.

When presenting the ToM measure, stimuli presentation matched in-person administration as closely as possible, with the exception that a gray circle was used to highlight relevant information (e.g., characters) rather than pointing, see Figure S1 for example. Participants tested remotely gave verbal responses, and if participants attempted to point at the screen the experimenter repeated the question and used their cursor to highlight response options when applicable. For example, for the story in Figure S1, an experimenter might repeat, "Where will Diana look for her snack – in the drawer? *Circle drawer with cursor* Or in the desk? *Circle desk with cursor, then remove cursor from participant's view*.

Figure S1

Example of Remote Testing Stimuli

Notes. Experimenter script is presented below each image.

Appendix S2. Methods: Measures for Validity Analyses

Theory of Mind Inventory-2 (TOMI-2, Hutchins et al., 2014). This measure was administered only to the Sacramento valley subsample. Parents reported whether 60 statements about social reasoning (e.g., "My child understands that people can lie to purposely mislead others.") were like or unalike their child. Parents reported using a sliding scale scored continuously from 0-20 points, with anchors at "Definitely Not", "Probably Not", "Undecided", "Probably", and "Definitely" like their child. We analyzed the TOMI-2 as the average of all items. If parents did not complete all items, data were scored as missing. Higher scores indicated better social reasoning, and the possible range of scores was 0-20 points.

Children's Social Understanding Scale (CSUS, Tahiroglu et al., 2014). This measure was administered only to the Sacramento valley subsample. Parents reported whether 42 statements (e.g., Recognizes that if a person wants something, that person will probably try to get it.") were like or unalike their child. Parents reported using a 4-point Likert scale in which 1: Definitely Untrue, 2: Somewhat Untrue, 3: Somewhat True, and 4: Definitely True. A total score in addition to six subscales were calculated as the average of all items and average of items within the scale respectively: Desire, Belief, Knowledge, Intention, Perception, and Emotion. If parents did not complete all items within the subscale, we scored that subscale as "missing" but included other completed scales. Higher scores indicated better social reasoning, and the possible range of subscales was 1-4.

Executive Function

Up/Down (Kramer et al., 2015). This measure was administered only to the Sacramento valley subsample. Participants were presented with black arrows on a white background pointing either up or down. After a training phase in which participants correctly identified which direction the arrows were pointing, they were told they would play a game in which they would

say “up” when the arrow was pointing down, and “down” when the arrow was pointing up. Participants completed practice trials until they were correct on four consecutive trials ($M_{\text{TrialsAdministered}}$ = 4.16, *SD* = 0.76), and were told, “That’s right!” if they were correct and were re-told the rules if they were incorrect. Participants then completed 20 trials without experimenter feedback. Verbal responses were coded offline, and participants received 0 points if they said the word that matched the arrow’s direction (e.g., “up” for an up arrow), 1 point if they first said the matching word but self-corrected (e.g., “up… no wait down!” for an up arrow), 2 points if they said a partial word before self-correcting (e.g., “uh… down!” for an up arrow), and 3 points if they only said the correct word (e.g., “down” for an up arrow). Each video was scored by two of seven trained research assistants scored, and discrepancies were resolved by a third coder. Reliability between the two coders who coded the largest proportion of participants (67% and 62% respectively) was excellent, ICC = .96, 95%CI [.91, .99], F(16, 16.9) = 55.70, $p <$ .001.

Happy/Sad (Lagattuta et al., 2011). This measure was administered only to the Sacramento valley subsample. Participants were presented with a picture of a cartoon happy face and sad face and correctly identified the emotions presented in these faces. Participants were then told that they would play a silly game, and when presented with a happy face should say “sad” and say “happy” when presented with a sad face. Participants repeated the rules after the experimenter, and then completed four practice trials with feedback from the experimenter. If the practice trial was correct, participants were told, “That’s right!,” and if the practice trial was incorrect participants were told the rules again for both trial types. If participants were incorrect on any of the practice trials, participants were administered four additional practice trials until all practice trials were correct ($M_{\text{TrialsAdministered}}$ = 4.17, *SD* = 0.81). Participants completed 20 test

trials without experimenter feedback. Each trial was scored as 0 (participant said the emotion that matched the face presented), 1 (participant said the full word of the emotion that matched the face presented, but corrected themselves, e.g., "sad, wait happy"), 2 (participant said a partial word, e.g., "ha-" or "s-" before self-correcting), or 3 (participant gave only the correct response). Therefore, participants could earn a score between 0 and 60. Each video was scored by two of six trained research assistants scored, and discrepancies were resolved by a third coder. Reliability between the two coders who coded the largest proportion of participants (56% and 73% respectively) was excellent, ICC = .97, 95%CI [.93, .98], $F(32, 32.3) = 54.20$, $p < .001$.

Dimensional Change Card Sort (DCCS, Zelazo, 2006). This measure was administered only to the Greater Boston subsample. Participants sorted test cards that varied by two dimensions: color (blue, red) and shape (boat, bunny). Participants were first instructed to sort based on color for the first 6 trials, and were then instructed to sort based on shape for the next 6 trials. Participants were reminded of the current rules before being handed each card, e.g., "Red ones go here, and blue ones go there. Here's a red boat, where does this go?" Participants then also received additional "border" trials. Participants were instructed to sort cards that either had a black border surrounding the picture or did not. Participants sorted black border cards by the 'color rule' and no border cards by the 'shape rule' for 6 trials, and sorted by the inverse for the final 6 trials. Participants had to sort 4/6 cards correctly in order to progress to the next sort rule. Scores on the DCCS were the sum of participants' correct trials, and therefore a composite of pre- and post-switch trials ranged from 0-12 points, and a composite of border trials also ranged from 0-12 points.

Kaufman Brief Intelligence Test 2 (KBIT-2, Kaufman & Kaufman, 2004). The verbal subscale of the KBIT-2 was administered to only the Sacramento valley subsample as a measure

of vocabulary. It was administered either in person using a picture book (n = 33) or by a web interface in which children clicked on their response (for younger children, children pointed to the screen and parents clicked on the child's response) (n = 64). Six images were presented on a page: the correct image and five distractor images. Participants continued the task until they gave an incorrect response for four consecutive pages/screens. For each item, participants pointed to or clicked on the item corresponding to the word or phrase read by the experimenter (e.g., "Point to/click on… clock"). Participants tested remotely completed a clicking training (in which they had to click on 4 items in a scene) and three practice items that followed a similar format to the KBIT-2, and then completed the KBIT-2. Raw scores were calculated as the last correct item minus the number of incorrect items, and raw scores were used in analyses.

Peabody Picture Vocabulary Test (Version 4). The PPVT-4 was administered to only the Boston subsample as a measure of vocabulary. Four images were presented on a page: the correct image and three distractor images. For each item, children pointed to the image corresponding to the word or phrase read by an experimenter. Children continued the task until they gave an incorrect response on four consecutive pages. Raw scores were calculated as the total number of correct items and raw scores were used in analyses.

Appendix S3. Preliminary Analyses of Group Differences and Missingness

To test whether ToM differed across groups, a composite was created as the average score of the items that participants completed. Most participants (61%) completed all items, and the remainder of participants completed 10 or more items.

Participants did not differ by gender based on geographic location, $\chi^2(2, N = 224) = 1.94$, $p = .379$. Participants from the greater Boston area were older ($M_{age} = 7.56$, $SD = 2.75$) than participants tested in the Sacramento Valley ($M_{age} = 6.66$, $SD = 1.10$), $t(174.02) = 3.36$, $p < .001$, and also scored higher on the ToM measure ($M = 0.81$, $SD = 0.16$) than children in the Sacramento Valley ($M = 0.74$, $SD = 0.13$). Within the Sacramento Valley sample, we also tested whether participants tested in the laboratory ($n = 33$) differed from participants tested remotely ($n = 64$). Participants tested remotely ($M_{age} = 7.02$, $SD = 0.92$) were older than participants tested in lab ($M_{age} = 5.97$, $SD = 1.09$), $t(55.9) = 4.74$, $p < .001$, and participants tested remotely trended toward higher ToM scores ($M = 0.76$, $SD = 0.12$) than participants tested in lab ($M = 0.71$, $SD = 0.13$), although differences were non-significant $t(61.31) = 1.90$, $p = .062$. Differences in ToM performance by geographic location and testing modality were likely due to differences in age, and would be expected given that ToM performance improves across age.

Next, we examined missing data patterns across items. There were minimal missing data in the ToM measure, such that across the 17 items analyzed in all samples, 11 items had no missing data, and remaining items had 1 – 33% missing. Missing data were handled using maximum likelihood for confirmatory factor and IRT analyses.

Appendix S4. Meeting IRT Assumptions

Achieving Local Independence

IRT assumes that data are locally independent, such that the probability of a correct response is dependent only on child parameters, such as the child's underlying ToM, and item parameters, such as the item's difficulty. Not meeting assumptions for local independence can bias parameter estimates, such that an item may appear unrelated to the latent factor because its variance is explained by a related item. Local independence was achieved by selecting one prediction question from each story within the ToM measure (Richardson et al., 2018), as described previously.

Assessing Model Unidimensionality in Exploratory and Confirmatory Factor Analysis

IRT assumes that there is a unidimensional structure of the latent factor, such that a single latent factor for ToM underlies children's responses. To test whether the measure met assumptions of unidimensionality, we first conducted an exploratory factor analysis to examine the best structural fit to the data, and then interpreted fit indices from a single-factor confirmatory factor analysis to confirm unidimensionality. Acceptable fit criteria were a $\chi 2$:df ratio ≤ 2 (Schreiber et al., 2006), Tucker-Lewis index (TLI) $\geq .90$ and comparative fit index (CFI) $\geq .90$ (Brown, 2006), and root mean square error of approximation (RMSEA) $\leq .08$ (Hu & Bentler, 1999).

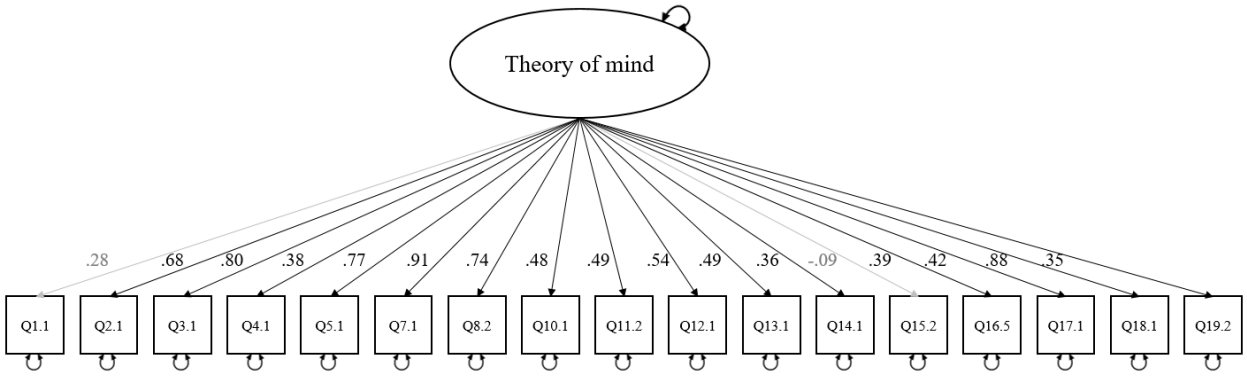
Scree plots of principal components and factor analysis decomposition eigenvalues suggested that a single-factor model best fit the data. Therefore, a unidimensional model for ToM was fit, in which the variance of the latent variable was fixed to 1 in order to scale parameters. The model was estimated using diagonally weighted least squares and responses were fit as ordinal indicators, and we report robust test statistics.

A unidimensional model for ToM had acceptable fit, in which the $\chi 2$:df ratio = 1.33, TLI = .91, CFI = .92, and RMSEA = .04, 95% CI [0.02, 0.05], see Figure S2 and Table S1. These fit indices are also in-line with prior investigations of the factor structure of ToM (Hughes et al., 2014; Devine & Hughes, 2013; Wang et al., 2016). Thus, the model met assumptions for unidimensionality.

This model's factor loadings ranged from -.09 to .91, see Figure S2. Factor loadings for all but one item (item #15.2) were positive, suggesting that better performance on any given item was related to better ToM understanding.

Figure S2

Confirmatory Factor Analysis for Meeting Unidimensionality Assumptions



Notes. The text for each item is available in Table 1 in the main text. Bolded factor loadings were significant, $p < .05$. Two factor loadings were non-significant (items #1.1 and #15.2).

Table S1. Fitted Covariance Matrix for Meeting Unidimensionality Assumptions of 17 Item CFA Model

	Q1.1	Q2.1	Q3.1	Q4.1	Q5.1	Q7.1	Q8.2	Q10.1	Q11.2	Q12.1	Q13.1	Q14.1	Q15.2	Q16.5	Q17.1	Q18.1	Q19.2
Q1.1	.92																
Q2.1	.19	.54															
Q3.1	.22	.54	.36														
Q4.1	.11	.26	.30	.86													
Q5.1	.22	.52	.62	.29	.40												
Q7.1	.25	.62	.73	.35	.70	.18											
Q8.2	.21	.50	.59	.28	.57	.67	.45										
Q10.1	.14	.33	.39	.18	.37	.44	.36	.77									
Q11.2	.14	.33	.39	.18	.37	.44	.36	.23	.76								
Q12.1	.15	.37	.43	.21	.42	.49	.40	.26	.26	.71							
Q13.1	.14	.34	.40	.19	.38	.45	.37	.24	.24	.27	.76						
Q14.1	.10	.24	.29	.14	.28	.33	.27	.17	.17	.19	.18	.87					
Q15.2	-.02	-.06	-.07	-.03	-.07	-.08	-.07	-.04	-.04	-.05	-.04	-.03	.99				
Q16.5	.11	.26	.31	.15	.30	.35	.29	.19	.19	.21	.19	.14	-.03	.85			
Q17.1	.12	.28	.33	.16	.32	.38	.31	.20	.20	.23	.21	.15	-.04	.16	.83		
Q18.1	.24	.59	.70	.33	.68	.79	.65	.42	.42	.47	.43	.32	-.08	.34	.37	.23	
Q19.2	.10	.24	.28	.13	.27	.32	.26	.17	.17	.19	.17	.13	-.03	.14	.15	.31	.88

Notes. The text for each item is available in Table 1 in the main text. Diagonal values represent the variance of each indicator, and off-diagonal elements represent the covariance between two indicators.

Table S2. Fitted Covariance Matrix for Final Unidimensional Factor Model

	Q1.1	Q2.1	Q3.1	Q4.1	Q5.1	Q7.1	Q8.2	Q10.1	Q11.2	Q12.1	Q13.1	Q14.1	Q16.5	Q17.1	Q18.1	Q19.2
Q1.1	.92															
Q2.1	.19	.54														
Q3.1	.22	.54	.36													
Q4.1	.11	.26	.30	.86												
Q5.1	.21	.52	.62	.29	.41											
Q7.1	.25	.61	.72	.34	.70	.18										
Q8.2	.20	.50	.59	.28	.57	.67	.46									
Q10.1	.13	.33	.39	.18	.37	.44	.36	.77								
Q11.2	.14	.33	.39	.18	.38	.44	.36	.24	.76							
Q12.1	.15	.38	.45	.21	.43	.50	.41	.27	.27	.69						
Q13.1	.14	.34	.40	.19	.38	.45	.37	.24	.24	.28	.75					
Q14.1	.10	.24	.29	.14	.28	.33	.27	.17	.18	.20	.18	.87				
Q16.5	.11	.26	.31	.15	.30	.35	.29	.19	.19	.22	.19	.14	.85			
Q17.1	.12	.29	.34	.16	.32	.38	.31	.20	.20	.23	.21	.15	.16	.82		
Q18.1	.24	.59	.70	.33	.68	.79	.65	.42	.43	.49	.43	.32	.34	.37	.23	
Q19.2	.10	.24	.28	.13	.27	.32	.26	.17	.17	.20	.18	.13	.14	.15	0.31	.87

Notes. The text for each item is available in Table 1 in the main text. Diagonal values represent the variance of each indicator, and off-diagonal elements represent the covariance between two indicators.

Table S3. Summary Statistics for Sacramento and Boston Samples

	Sacramento		Boston		Possible Range
	N	*M* (SD)	*N*	*M* (SD)	
Age (years)	97	6.66 (1.10)	127	7.56 (2.75)	3 – 12
Vocabulary					
KBIT-2	92	23.39 (5.92)	-	-	0 – 60
PPVT-4	-	-	93	162.69 (27.27)	20 – 160
ToM					
ToM Average Score	97	0.76 (0.13)	127	0.82 (0.17)	0 – 1.00
TOMI-2: Total	77	16.79 (1.77)	-	-	0 – 20
TOMI-2: Early	85	18.05 (1.33)	-	-	0 – 20
TOMI-2: Basic	86	17.83 (1.49)	-	-	0 – 20
TOMI-2: Advanced	86	15.08 (2.59)	-	-	0 – 20
TOMI-2: Emotion	88	17.25 (1.79)	-	-	0 – 20
TOMI-2: Mental State	89	18.35 (1.71)	-	-	0 – 20
TOMI-2: Pragmatics	88	15.93 (2.72)	-	-	0 – 20
CSUS: Total	61	3.36 (0.28)	-	-	1 – 4
CSUS: Desire	87	3.53 (0.38)	-	-	1 – 4
CSUS: Knowledge	82	3.61 (0.37)	-	-	1 – 4
CSUS: Belief	82	3.48 (0.41)	-	-	1 – 4
CSUS: Emotion	80	3.19 (0.34)	-	-	1 – 4
CSUS: Intention	83	3.40 (0.35)	-	-	1 – 4
CSUS: Perception	74	2.85 (0.37)	-	-	1 – 4
Executive function					
Up/Down	61	56.13 (6.60)	-	-	0 – 60
Happy/Sad	93	54.17 (7.61)	-	-	0 – 60
DCCS: Pre- & post-sw.	-	-	31	10.77 (1.98)	0 – 12
DCCS: Border	-	-	25	6.84 (1.77)	0 – 12

Notes. Dashes indicate that the measure was not collected in that sample. KBIT-2: Kaufman Brief Intelligence Test-2, Vocabulary subscale. PPVT: Peabody Picture Vocabulary Test. ToM

average score was the average of items that each participants completed in order to maximize the number of participants included in validity analyses. TOMI-2: Theory of Mind Inventory-2. CSUS: Children's Social Understanding Scale. DCCS Pre & Post-sw: Dimensional Change Card Sort, Pre- and Post-switch trials (color and shape game). DCCS: Border: Dimensional Change Card Sort, Border trials.

Table S4. Relational Validity Correlations within the Sacramento Sample

		1	2	3	4
1	ToM	-			
2	Age (years)	0.46***	-		
3	KBIT-2	0.51***	0.69***	-	
4	Up/Down	-0.01	0.24	0.34**	-
5	Happy/Sad	0.07	0.21*	0.28**	0.18

Notes. Pairwise correlation coefficients (*r*). ***$p < .001$; **$p < .01$; *$p < .05$.

Table S5. Relational Validity Correlations within the Boston Sample

		1	2	3	4
1	ToM	-			
2	Age (years)	0.74***	-		
3	PPVT-4	0.50***	0.82***	-	
4	DCCS: Pre- & post-switch	0.39*	0.26	-	-
5	DCCS: Border trials	0.17	0.00	-	0.03

Notes. Pairwise correlation coefficients (*r*). Correlations between the PPVT-4 and DCCS could not be estimated because no participant completed both tasks. ***$p < .001$; **$p < .01$; *$p < .05$.

Chapter 2: Comprehensive assessment of theory of mind (CAT): A novel assessment of 3- to 8-year-old children's theory of mind and evaluation of mental-state scaling

Abstract

Theory of mind (ToM) describes the understanding that mental states guide behavior in the real world, and is a broad construct that is often assessed through measures that include multiple mental states (e.g., Wellman & Liu, 2004). However, existing measures fail to capture the richness and complexity of children's ToM given limitations in the measurements themselves, such as a focus on one test question type (i.e., 'prediction' questions) and the inclusion of only one task to represent each mental state concept. These limitations have made it challenging to fully characterize the developmental trajectory of ToM, and to examine how variability in task structure and domain-general demands influence the scaling of mental states. The present study developed a novel multi-mental state measure of ToM—referred to as the Comprehensive Assessment of ToM (CAT)—that includes 3-6 stories each about diverse desires, diverse beliefs, knowledge access, knowledge expertise, false belief, and visual perspective taking, as well as a non-social representational measure (e.g., false sign). All stories followed a similar structure, and included a prediction, explanation, and general comprehension question. With this measure, we replicate prior findings in which children more easily reason about desires compared to knowledge or false beliefs in a large sample (n = 206). We also extend prior findings demonstrating that when administering repeated measures of mental states and multiple question-types, aspects of knowledge, beliefs, and false beliefs develop alongside each other, and alongside visual perspective-taking and false signs. Taken together, our novel ToM assessment reveals a more nuanced pattern of difficulty scaling that extends over early to middle childhood.

Theory of mind (ToM) describes the understanding that internal mental states, including intentions, desires, thoughts, knowledge, and emotions, guide human behavior in the real world. Clear ToM developments occur across the preschool years when children come to understand mental states as person-specific representations of the world that can be false, as typically indicated by a transition from failing to passing 'false-belief tasks' in which children predict the actions of a character whose belief (e.g., about an object's location) has become false (Wellman et al., 2001). Prior to passing false-belief tasks, children appear to progress in their understanding of other mental states. For example, toddlers can explicitly reason about their own and others' desires, and in early preschool they begin to understand others' knowledge states (as knowledgeable or ignorant) and beliefs (as independent from reality). Indeed, ToM as a field was revolutionized by foundational work by Wellman and Liu (2004) showing that children all over the world progress in their mental-state understanding from first being able to accurately reason about desires, then about beliefs/knowledge, then false belief, then hidden emotions (Liu et al., 2008; Peterson et al., 2012; Shahaeian et al., 2011; Wellman & Liu, 2004; Wellman et al., 2006; 2011). This multi-mental state measure dominates the field of ToM research in preschool: In a review of comprehensive ToM measures by Beaudoin and colleagues (2020), over half (65%) of studies employed the ToM scale developed by Wellman and Liu (2004), whereas the remaining measures (e.g., ToM storybooks, Blijd-Hoogewys et al., 2008; the ToM test, Muris et al., 1999, and 11 other ToM measures) were split across the remaining 35% of studies.

Despite the incredible value the Wellman and Liu (2004) ToM scale brings to the study of ToM, it is limited in its ability to reveal a full and nuanced characterization of ToM and its development. First, the original Wellman and Liu scale consists of a singular task item to represent each mental-state construct. In contrast, other fields of cognition administer repeated

trials/items assessing a single construct (e.g., executive function measures, grass/sky, Carlson & Moses, 2001) even when assessing young infants (e.g., attention measures, IOWA cueing task, Ross-Sheehy et al., 2015), because the addition of more trials decreases measurement error (Spearman, 1910). Thus, children's performance over multiple items of a given mental-state construct (e.g., desires) may reveal a more nuanced difficulty in relation to other assessed constructs (e.g., knowledge, beliefs).

Second, the original scale leaves out additional relevant constructs that may be critical to ToM development. In particular, there are both theoretical and empirical connections between ToM and visual perspective taking—the understanding of the connection between gaze/visual access and mental state content (e.g., *knowing* the contents of a box requires *looking* inside). Early ToM research has shown that children develop a rudimentary understanding of visual perspective in toddlerhood (Flavell, 1978), although children develop in their understanding of more complex perspective taking through preschool (e.g., Carlson et al., 2004), middle childhood (e.g., Richardson et al., 2018), and even show variability in adulthood (e.g., Todd et al., 2017). Visual perspective taking has also been proposed as a construct that is a part or a process within ToM (Apperly, 2012; Frith & Frith, 2006; Gerrans & Stone, 2008). Thus, visual perspective taking remains a fundamental component of mental-state reasoning throughout the lifespan, but has not been scaled along with the traditional ToM measures assessed in preschool (e.g., in Wellman and Liu, 2004). Similarly, there are both theoretical (Carey, 1985; Gopnik & Wellman, 1992; 1994) and empirical connections (Iao et al., 2011; Leekam et al., 2008; Sabbagh et al., 2006) between non-social representational reasoning (e.g., false photograph, false sign) and ToM (particularly false belief). For example, the false sign task (Parkin, 1994) follows the same story structure as false belief with the exception that rather than a *belief* about an object's location

becoming false, a pictorial representation (e.g., a signpost, an arrow) changes so that it no longer reflects reality. Thus, false sign is similar to false belief in that both tasks assess an understanding of representational change and false representations (Sabbagh et al., 2006), and domain-general abilities (e.g., matched in vocabulary and executive function; Iao et al., 2011). Thus, open questions remain about how mental-state constructs such as desires, beliefs, and knowledge may relate to these additional constructs and how they 'scale' alongside them.

Some existing studies have expanded the original scale by adding measures of other concepts that require mental-state reasoning (i.e., sarcasm, Peterson et al., 2012), while other more recent studies have scaled altogether different mental-state items (e.g., first-, second-, and third-order false belief, morally relevant false belief, faux pas, strange stories; Osterhaus et al., 2016; 2022). However, similar limitations of few or singular items to represent a given mental-state construct apply to these more recent scales as well. Moreover, a third limitation exists in these and the original Wellman and Liu (2004) scales, in that the general task demands and 'story structure' of the individual task items are not equated (e.g., some have lengthy story set-ups with multiple control questions (e.g., hidden emotion), whereas others are relatively straightforward (e.g., explicit false belief, in which children are told what a character *thinks* and then asked immediately after how the character will act), and others still have memory checks or engagement questions before the final 'test' prediction item, e.g., in a knowledge access task, children are asked what is in the box after they are shown). This leaves open the possibility that different mental-state items scale on the basis of difficulty in general task demands, or may be skewed by over- or under-representation in the scale overall.

Finally, existing studies have constrained their scaling analysis to prediction questions that require participants to predict a characters' actions on the basis of their mental state.

However, a few more recently-developed assessments of ToM include questions in which participants must reason about *why* a character does a certain action. These 'explanation questions' offer an additional means of assessing ToM, and potentially tap a richer understanding of children's mental-state reasoning, although they remain less-frequently assessed in the field and have not been examined psychometrically. Thus, open questions remain about the role of explicit and post-hoc reasoning about actions based on mental states, how explanation question difficulty scales in relation to prediction questions alone, and whether explanation questions do indeed tap a richer, more complex understanding of the mind or whether they add measurement noise when combined with prediction questions alone.

Thus, there remain unanswered questions about ToM measurement and the difficulty scaling of mental states. Specifically, it is not known whether the documented mental-state scaling in Wellman and Liu (2004) remains consistent when there are multiple assessments of each mental state, when ToM measures include both prediction and explanation questions, and when general task demands and 'story structure' are better equated across items. Moreover, it is unknown how additional constructs hypothesized as foundational to multiple mental states (e.g. visual perspective-taking, false sign, Carlson et al., 2004; Flavell, 1978; Parkin, 1994) may scale alongside highly-studied constructs of desires, diverse beliefs, knowledge, and false beliefs. Additionally, no study has examined difficulty scaling of different item contexts (e.g., whether participants are asked to compare their own mental state versus another, or compare mental states across two different characters), or different aspects of mental-state concepts (e.g., understanding access to knowledge as well as knowledge expertise). New investigations are needed to clarify and reveal potential nuances in how multiple mental-state constructs scale in difficulty across childhood. More broadly, given predominant use of the original Wellman and

Liu (2004) scale in the field (see Beaudoin et al., 2020), much of what we know about ToM development more generally over early childhood is constrained by the use of the scale and its limitations. Thus, novel measures of ToM are needed that address the above-mentioned limitations to advance our understanding of how different mental states scale over childhood.

The Present Study

The present study aims to develop a novel measure of children's ToM that addresses the limitations of existing ToM scales. Namely, we have developed the Comprehensive Assessment of ToM (CAT) which expands the original Wellman & Liu (2004) scale in several key ways. First, in addition to assessing mental states most commonly administered within the child ToM literature including diverse desires, diverse beliefs, knowledge access, and false belief, the CAT also assesses children's understanding of knowledge expertise, visual perspective taking and non-social representations (i.e., false signs). Second, the CAT includes 3-6 items to assess each mental-state category in order to assess the robustness of difficulty scaling when there are additional items, and to examine how variability in question format influences scaling. Third, each item is matched in terms of its structure or format, such that participants are first told a brief story to set-up important context, asked a prediction question, followed by an explanation question, and then finally a control question to assess general story comprehension. This consistent task structure matches difficulty due to response demand (e.g., an explanation response involves a greater verbal demand compared to a prediction) in order to reveal variability in difficulty due to mental-state content. Fourth, we administer the CAT to a wide age range of children from 3- to 8-years-old, in order to shed light on the developmental trajectory of ToM beyond the preschool years. We conduct psychometric analyses of this novel ToM assessment in an item response theory framework (IRT, see Embretson & Reise, 2000) in a

large-scale (n = 206) sample of children in order to examine the developmental trajectory of different mental states and false sign, and to assess reliability of the CAT. In sum, the present study introduces a novel multi-mental state assessment of ToM in a large sample in order to reveal a more nuanced and comprehensive analysis of mental state development over the preschool years and beyond.

Methods

Participants

A large sample (n = 206) of 3- to 8-year-old children were tested remotely via Zoom. Participants were recruited from a database of families willing to participate in research, and compensated for their time with a $15 giftcard. The Institutional Review Board approved all methods and procedures used in this study, and all parents gave informed consent prior to participation. Seventeen subjects were excluded from the final sample: 11 participants did not meet a minimum vocabulary score required to understand and respond to the theory of mind task, 4 were excluded due to technical issues with presenting stimuli (e.g., internet speed), 1 due to technical errors in recording videos for later data entry, 1 because the participant did not complete both testing sessions, and 1 child refused to participate. We recruited an approximately even number of boys and girls across each age group, although 3-year-olds were more likely to be excluded due to low vocabulary score. The final sample consisted of 26 3-year-olds (10 girls, 16 boys, M_{age} = 3.67 years, SD = 0.20), 34 4-year-olds (17 girls, 17 boys, M_{age} = 4.54, SD = 0.27), 35 5-year-olds (17 girls, 19 boys, M_{age} = 5.54, SD = 0.19), 36 6-year-olds (17 girls, 18 boys, M_{age} = 6.63 years, SD = 0.15), 35 7-year-olds (18 girls, 16 boys, and 1 genderfluid child, M_{age} = 7.60 years, SD = 0.15), and 40 8-year-olds (21 girls, 19 boys, M_{age} = 8.65 years, SD = 0.11). Participants' demographic makeup reflected the diversity of the Sacramento Valley: 96 were White (19% Latinx, Chicanx or Hispanic), 38 were multi-racial (44% Latinx, Chicanx or

Hispanic), 13 were Asian or Asian- American (8% Latinx, Chicanx or Hispanic), 4 were African or African American (not Latinx, Chicanx or Hispanic), 1 was American Indian or Alaskan Native (Latinx, Chicanx or Hispanic), 5 reported race as "Other" (100% Latinx, Chicanx or Hispanic), 3 subjects did not report race (100% Latinx, Chicanx or Hispanic), and 46 subjects did not report race nor ethnicity. The median educational attainment of the child's primary caregiver was a four-year college degree (n = 49); 1 completed 8th grade, 4 had a high school diploma or equivalent, 16 had some college education, 16 had an Associate's or technical degree, 12 had some graduate education, 44 had a Master's degree, 15 had a Ph.D. or M.D., 3 reported education as "Other", and 46 did not report education. The median educational attainment of the child's secondary caregiver was a four-year college degree (n = 53); 1 completed 8th grade, 2 completed some high school, 9 had a high school diploma or equivalent, 25 had some college education, 17 had an Associate's or technical degree, 5 had some graduate education, 23 had a Master's degree, 13 had a Ph.D. or M.D., 4 reported education as "Other", and 54 did not report education. Median family income was $100,000 and greater ($n$ = 100); 1 earned $15-24k, 14 earned $25-49k, 5 earned $50-74k, 30 earned $75-100k, and 61 did not report income.

Measures

Comprehensive Assessment of Theory of Mind (CAT)

Participants competed at least three items and up to six items within each mental state category (i.e., desires, knowledge, visual perspective taking, discrepant and true beliefs, and false belief), and false sign (i.e., a non-social control condition for false belief). For each item, participants viewed an animated story, and were asked several questions to assess their understanding of mental states. For each story, there was a prediction question (in which the participant had to predict a character's action). Then, the character acted in accordance with their

mental state. All participant (regardless of whether their prediction was correct or incorrect) were then asked to explain why the character performed that action. This explanation was then coded for whether the response was correct or incorrect. Correct responses had to implicitly or explicitly reference the mental state possessed by the character, See explanation coding scheme at https://osf.io/4d9ar. At the end of each story, there was a non-social control question about relevant story information to account for performance based on non-social skills such as attention, and to ensure children understood the critical content of the story. Items were scored such that if participants failed the control question, they received a score of 0 for the item overall in order to ensure that participants' responses reflected an understanding of the story context, and following scoring of similar story-based ToM tasks (Peterson et al., 2012; Wellman & Liu, 2004; Wellman et al., 2011). The materials and full coding scheme are available at https://osf.io/4d9ar.

Diverse desires. Diverse desires was assessed through three items in which the character's preference differed from the participant's (following the format of Repacholi & Gopnik, 1997 and Wellman & Liu, 2004) and three items in which the participant was told about a boy's and girl's preferences. Participants were then asked who would want the item/snack revealed (following the format of Bowman et al., 2012). Predictions were scored as correct if the participant predicted the character would act in accordance with the character's preference rather than the participant's own preference or the other character's preference. Explanations were correct if the participant referenced the character's preference (e.g., "He wanted grapes," "he likes grapes"). There were a total of six items that measured diverse desires.

Diverse beliefs. Diverse beliefs was assessed through two items in which the character's belief differed from the participant's (following the format of Gopnik et al., 1994 & Wellman & Liu, 2004) and two items in which the participant was told about a boy and girl's beliefs and then

asked who was correct when the item/snack's location was revealed (following the format of Bowman et al., 2012 and Liu et al., 2009). Predictions were scored as correct if the participant predicted the character would act in accordance with the character's belief rather than the participant's own belief or the other character's belief. Explanations were correct if the participant referenced the character's belief (e.g., "She thought the cat was hiding in the garage.") There were a total of four items that measured diverse beliefs.

Knowledge expertise. For each item, participants were presented with two characters introduced as having vocations that children would find familiar (e.g., a chef and a pilot). Participants were then asked about the relative expertise of each character (e.g., "Who knows more about how fast a plane needs to go before it can take off?", adapted from Lane et al., 2014). Predictions were correct if they identified the more knowledgeable character for each scenario (e.g., "the pilot knows more"). Explanations were correct if participants identified that a character's follow-up actions or statements were based on that character having superior knowledge in their relevant domain (e.g., "the pilot flies a lot of planes"). There were three items (with pairings across a mechanic, chef, and pilot) that measured knowledge expertise.

Knowledge access. Participants were presented with a container (e.g., a box) and asked what was inside. The contents were revealed to the participant (e.g., a flower). Participants were presented with a novel character and told that this character had never seen inside of the container (following task formats in Pratt & Bryant, 1990; Pillow, 1989; Wellman & Liu, 2004). Predictions were correct if participants said that the character would not know what was inside, and explanations were correct if they referenced the character's lack of visual access or knowledge (e.g., "She didn't see inside," "he doesn't know what's inside"). There were a total of three items that measured knowledge access.

Visual perspective taking. Visual perspective taking assessed participants' understanding that people can have different visual perspectives, and that someone will be thinking about the object that they are looking at rather than one they cannot see. Participants were presented with stories in which one character's vision was occluded (e.g., a hat was covering their eyes) while another character had clear visual access to a target object/scene (similar to tasks in Carlson et al., 2004; Piaget & Inhedler, 1969; Richardson et al., 2018). Participants were asked which character had unoccluded visual access (e.g., who to ask about the color of a bird in the sky) or to predict a character's action based on their occluded visual access (e.g., which object a character will reach for). Predictions were correct if they predicted a character's action based on their occluded visual access or if they could identify the character with unoccluded visual access. Explanations were correct if they referenced a character's occluded vision or a character's lack of knowledge based on lack of visual access. There were a total of six items that measured visual perspective taking.

False belief. Three versions of a false-belief 'location-change' task and three versions of a false-belief 'unexpected contents' task were administered (similar to tasks in Baron-Cohen et al., 1985; Perner et al., 1989; Wimmer & Perner, 1983; Wellman & Liu, 2004). In the location change tasks, one character moved an item while another character was away from the scene (e.g., Ryan moved a cookie from the oven into the refrigerator while Dan was taking a walk). Predictions were correct if participants stated that the character would search in the location where they last saw their item. Explanations were correct if participants referenced that a character's actions were motivated by a lack of knowledge or lack of visual access to the object being moved, or by the character's thought/belief that the object was in that specific location (e.g., "He doesn't know the cookie was moved"). In the unexpected contents task, participants

were shown a picture of a prototypical container (e.g., a Band-Aids box, a crayon box) that contained an unexpected item (e.g., marbles, candles). Children's predictions were correct if they stated that a novel character would believe that the contents matched the box (e.g., that there were band-aids in the band-aids box), and explanations were correct if they referenced the character's lack of knowledge or visual access to the unexpected contents, or the character's thoughts/beliefs about the contents based on the box (e.g., "He doesn't know there's really marbles inside," "he thinks it because there's Band-Aids on the box"). There were a total of six items that measured false belief.

True belief. True belief understanding was assessed with tasks following the same format as the false-belief location-change stories, except that the target character observed their item being moved (following tasks in Fabricius & Khalil, 2003; Fabricius et al., 2010; Richardson et al., 2018). Predictions were correct if participants stated that the character would search in the item's current location where they had observed it be moved to, and explanations were correct if they referenced that the character knew or saw that the item had been moved. There were a total of three items that measured true belief. Note that due to poor psychometric properties (see Results section below), these items were not included in focal age-related or difficulty scaling analyses.

False sign. In the false sign items, a sign (e.g., sign post, sticker) indicates where an object or character is located, but the sign changes directions (e.g., the wind blows the sign to face another direction, following tasks in Parkin, 1994; Ford et al., 2012). Predictions were correct if participants stated that a character would search in accordance with the sign rather than reality, and explanations were correct if participants referenced that the character followed the

sign or thought the item would be in the location the sign pointed to/represented. There were a total of six items that measured false sign.

Executive Function

Grass/sky (Carlson & Moses, 2001). Participants completed a conflict-inhibition stroop task that assessed their ability to inhibit salient perceptual features that shared semantic similarity in order to respond to a non-dominant feature. Participants were instructed to touch a picture of a blue sky with clouds when they heard an audio recording of the word "grass", and to touch a picture of grass when they heard "sky", using a touchscreen monitor. There were six practice trials and sixteen test trials. Trials were presented semi-randomized such that there were no more than two trials of the same type consecutively (e.g., grass, grass, sky). Participants had to respond correctly on at least two practice trials (one of each trial type – grass and sky) to be included in analyses, per Carlson and Moses, 2001). Participants received one point per correct response for each of 16 test trials.

Up/down (Kramer et al., 2015). Participants were presented with black arrows on white background pointing either up or down. After a training phase in which participants correctly identified which direction the arrows were pointing, participants were told that they would play a game in which they would say "up" when the arrow was pointing down, and "down" when the arrow was pointing up. Participants completed practice trials until they were correct on four consecutive trials ($M_{\text{Number of practice trials administered}} = 4.44$, $SD = 1.25$), and were told, "That's right!" if they were correct and were re-told the rules if they were incorrect. Participants then completed 20 trials without experimenter feedback. Verbal responses were coded offline, and participants received 0 points if they said the word that matched the arrow's direction (e.g., "up" for an up arrow), 1 point if they first said the matching word but self-corrected (e.g., "up… no wait down!"

for an up arrow), 2 points if they said a partial word before self-correcting (e.g., "u--… down!" for an up arrow), and 3 points if they only said the correct word (e.g., "down" for an up arrow). Two coders scored each child's video, and a third coder determined the correct code for trials with discrepancies. Reliability was calculated between two coders who coded a significant proportion of data (approximately 30% of the total data) before a third coder resolved discrepancies, and was high, ICC = .99, CI_{95} [.99, 1.00].

Vocabulary

Kaufman Brief Intelligence Test 2 (KBIT-2, Kaufman & Kaufman, 2004). The verbal subscale of the KBIT-2 was administered as a measure of children's vocabulary and was administered by a web interface in which children clicked on their response (for younger children, children pointed to the screen and parents clicked on the child's response). Six images were presented on a page: the correct image and five distractor images. Children continued the task until they gave an incorrect response for four consecutive pages. For each item, children clicked on the item corresponding to the word or phrase read by the experimenter (e.g., "Click on… clock," "Click on… what lives in a forest"). All children completed a clicking training (in which they had to click on 4 items in a scene) and three practice items that followed a similar format to the KBIT-2, and then completed the KBIT-2 until they were incorrect on four consecutive items. Raw scores were calculated as the last correct item minus the number of incorrect items, and raw scores were used in analyses.

Procedure

Children ages 4- to 8-years-old participated in a one two-hour session, while 3-year-old children participated in two one-hour sessions approximately one-week apart (*M* = 6.14 days, *SD* = 5.68) to prevent fatigue. ToM items were divided into three blocks, and block order was

counterbalanced across participants. Tasks were administered in the following order: Block 1 of ToM, Up/Down, Block 2 of ToM, Block 3 of ToM, KBIT-2, and Grass/Sky. Participants engaged in movement-based break activities between each task.

Within each block and within each mental state, items were counterbalanced and presented in a fixed order. There were six different counterbalance orders for the ToM block presentation (i.e., possible combinations for presentation of the three blocks were 123, 132, 213, 231, 321, and 312). Mental-state stories were counterbalanced across the salience of response options: for forced-choice prediction questions, the experimenter presented the correct choice option first for half of the questions, and the correct choice option second for the remaining questions. For stories that contained more than one character, the character presented first was associated with the correct answer for half of the questions; for the other half, the character presented last was associated with the correct answer. The positioning of the correct character or object within each set was also counterbalanced wherein for half of the stories, the correct choice appeared on the right side of the screen, and in the other half it appeared on the left side. Finally, within each block, in half of the questions the boy was associated with the correct answer, and in the other half the girl was associated with the correct answer. Within each ToM block, stories corresponding to mental state categories were semi-randomized such that there were an equal number of stories about desires, knowledge, beliefs, etc., and no two stories assessing the same mental state were presented consecutively.

KBIT-2 and Grass/sky were presented by a web-based application (Gorilla, Anwyl-Irvine et al., 2020), and of the subjects who participated in these tasks, most subjects responded via touchscreen or computer (n = 172), some children responded with parent assistance (n = 2), and

some parents or other family member responded by clicking what the child pointed to on the device (n = 24).

Results

Data Analysis Plan

The present study evaluated difficulty scaling and item discrimination of a novel measure of children's ToM, the Comprehensive Assessment of ToM (CAT) in an item response theory (IRT) framework.

We first examined relations across mental state constructs and false sign with age (e.g., looking for expected improvement on items as children age), vocabulary and executive function, given expected relations between ToM and these related constructs (Devine & Hughes, 2014; Milligan et al., 2007; Wellman & Liu, 2004). We next fit two IRT models in order to examine the difficulty scaling of mental state constructs and false sign. The first model more directly followed existing scaling efforts (Osterhaus et al., 2016; 2022, Shahaeian et al., 2011; Wellman & Liu, 2004) in that only prediction items (and not explanation items) were included in the model. The second model used a sum score of prediction and explanation items (i.e., each item was scored from 0-2 in which children could earn 1 point each for prediction and for explanation) to examine a more comprehensive scaling of mental states and false sign. We draw comparisons across these two models to inform the field's understanding of the developmental trajectory of ToM.

Data analyses were conducted in R (Version 3.6.2; R Core Team, 2019). IRT models were fit using the ltm (Version 1.1-1, Rizopoulos, 2006) and mirt (Version 1.37.1, Chalmers, 2012) packages.

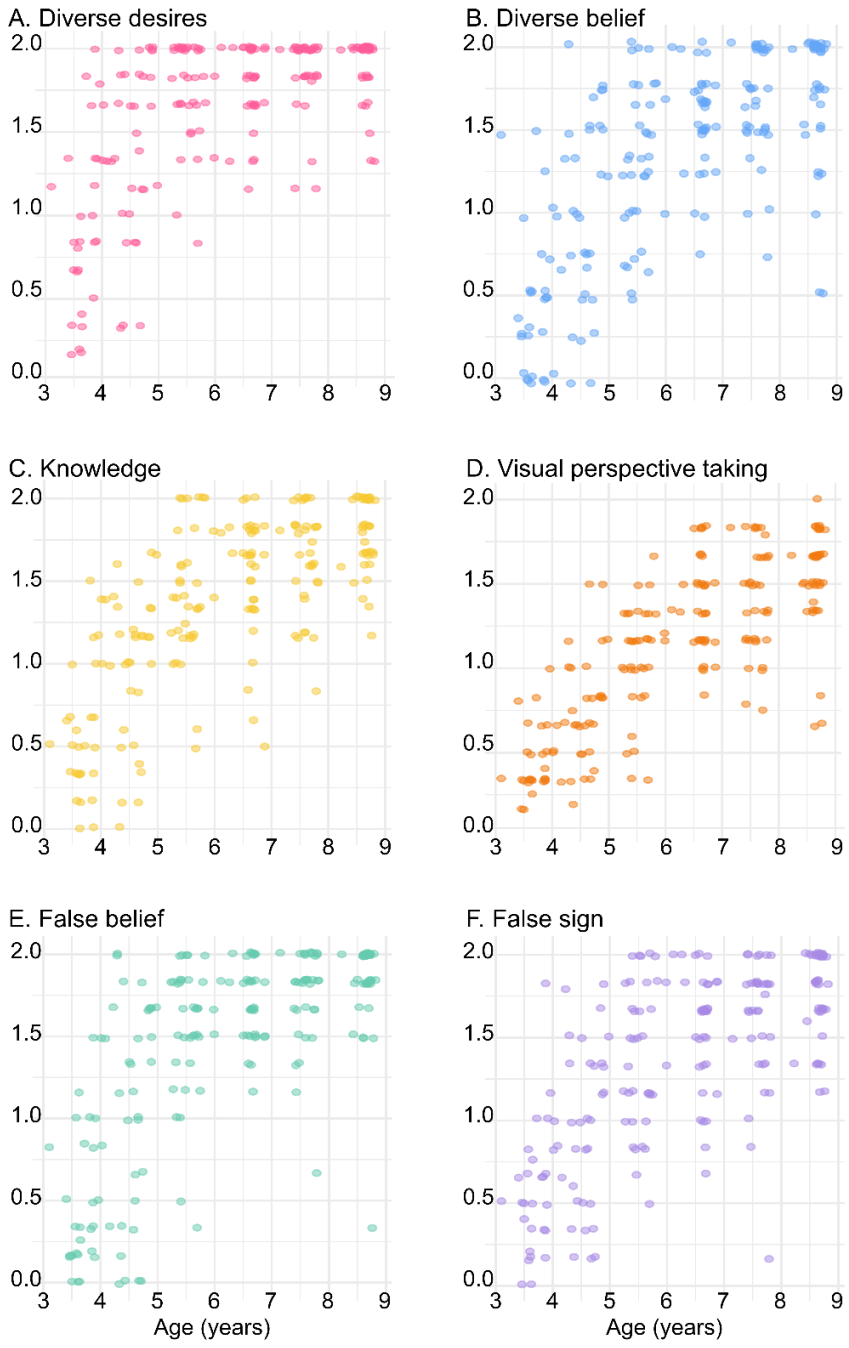
Preliminary Results: Examining the Comprehensive Assessment of ToM (CAT)

First, we conducted analyses to evaluate our novel assessment of children's ToM. All items had good psychometric properties (see details below) with one exception: True belief items in our sample showed the lowest reliability (Cronbach's $\alpha = 0.55$, whereas for all other mental states α was between 0.76 and 0.88) and poor discrimination (in which the three true belief items had the first, second, and fourth worst/lowest discrimination). Thus, true belief items were removed from the scale and subsequent analyses do not include them.

In order to conduct preliminary analyses to examine effects of age and relations among mental states and false sign, we formed composites of each mental state category and of false sign by taking the average performance of the participant's completed items in each category, and therefore scores were scaled the same as individual items from 0 to 2. We observed age-related improvements across items and mental-states as with age, rs > 0.60, ps < .001, see Figures 1 and 2. In addition, we also saw differences in overall performance across mental states, in which children began to pass desires (see Figure 2, in pink) before other mental states and false sign. We present item-level and average performance across ToM and false sign in Supplemental Materials, see Tables S1 and S2.

Figure 1

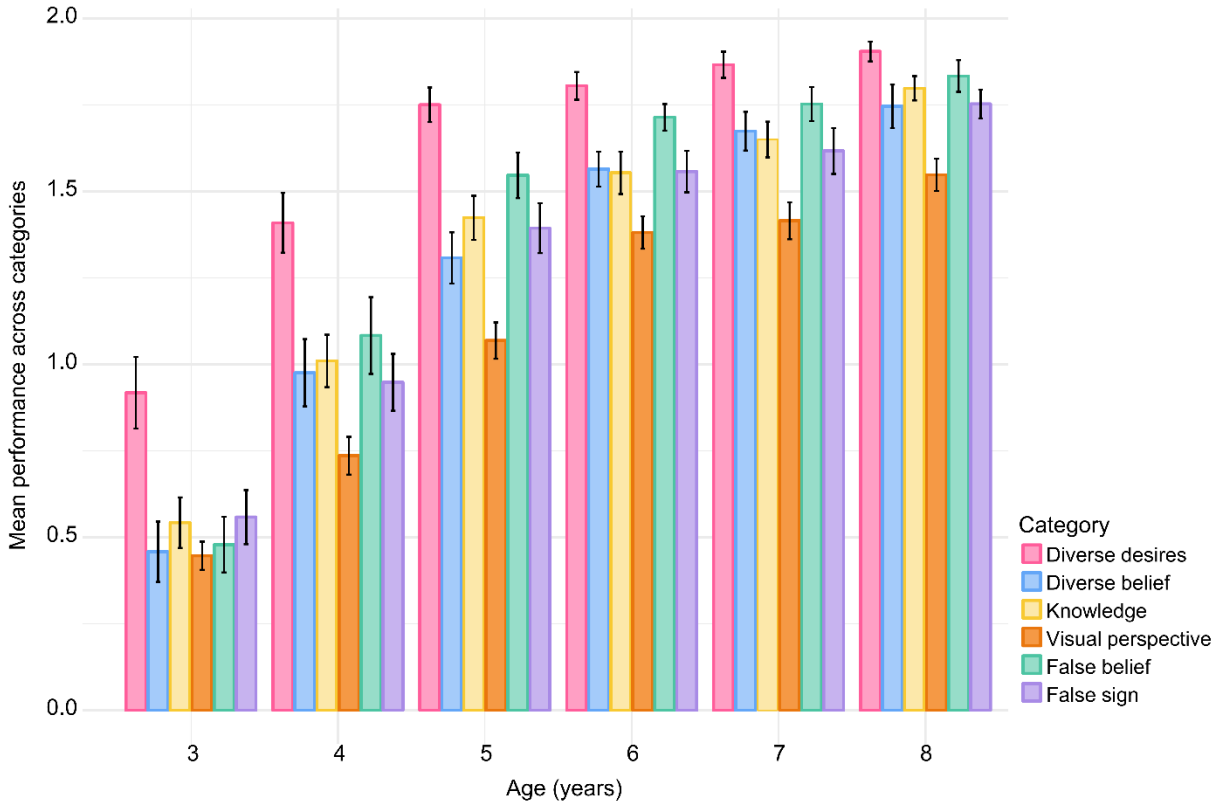
Scatterplots Illustrating Improvement in Mental-state Constructs and False Sign with Age



Note. Scatterplots depict the relation between participants' age and their average performance across mental states and false sign (from 0-2, in which 2 represents that participants got both prediction and explanation questions correct within the category).

Figure 2

Histogram Illustrating Improvement in Mental-state Constructs and False Sign with Age



Note. Histogram showing age-related improvements in children's ToM and false sign understanding (from 0-2, in which 2 represents that participants got both prediction and explanation questions correct within the category).

To examine relations between mental-state performance and domain-general constructs (e.g., age, vocabulary, and executive function), we examined bivariate and partial correlations (controlling for age) between a composite of each mental-state category, false sign, age, vocabulary, and executive function, see Table 1. In bivariate correlations, mental-state composites were all correlated with false sign, *r*s > 0.71, *p*s < .001, age, *r*s > 0.60, *p*s < .001, vocabulary, *r*s > 0.49, *p*s < .001, and executive function, *r*s > 0.23, *p*s < .003. In partial

correlations controlling for age, some of these relations dropped below significance. Specifically, there were weaker relations between EF and mental states, in which only diverse beliefs and visual perspective taking correlated with one of the EF measures (Grass/sky), and weaker relations between vocabulary and mental states, in which only knowledge and visual perspective taking correlated with vocabulary. There remained strong relations across mental states and between mental state reasoning and false sign.

Table 1

Bivariate and Partial Correlations between Children's ToM and Domain-General Skills

		1	2	3	4	5	6	7	8	9	10
1	Age	-	-	-	-	-	-	-	-	-	-
2	Vocabulary	.81***	-	.04	-.03	.24***	.22**	.05	.12	.14	.05
3	Diverse desires	.60***	.49***	-	.50***	.49***	.42***	.68***	.51***	.07	.11
4	Diverse beliefs	.68***	.54***	.70***	-	.43***	.42***	.60***	.48***	.20*	.06
5	Knowledge	.71***	.69***	.71***	.70***	-	.52***	.57***	.61***	.08	.06
6	Vis perspective	.77***	.71***	.68***	.72***	.78***	-	.46***	.51***	.20*	.04
7	False belief	.67***	.57***	.81***	.78***	.78***	.73***	-	.60***	.15	.03
8	False sign	.68***	.61***	.71***	.72***	.79***	.76***	.78***	-	.11	.07
9	Grass/sky	.40***	.41***	.29**	.40***	.34***	.43***	.38***	.35***	-	.08
10	Up/down	.33***	.32***	.26***	.25***	.26***	.27***	.23**	.26***	.18	-

Notes. Vis perspective = Visual perspective taking. Bivariate correlations are shown on the lower diagonal, and partial correlations controlling for age are shown on the upper diagonal. Pairwise deletion was used to conduct correlations when data were missing. * $p<.05$, ** $p<.01$, ***$p<.001$.

Next, we examined internal consistency for the ToM measure as whole, and within each mental state. Consistency was high for all subscales of ToM and for false sign, and ranged from Cronbach's $\alpha = .76$ to .96, see Table S3 in Supplemental Materials. Results supported that our novel measure of ToM was both related to constructs relevant to ToM development (e.g., age, vocabulary, executive function), and had high internal consistency.

Scaling of Mental States

In order to meet model assumptions for IRT analyses, data must be locally independent and unidimensional. We evaluated unidimensionality by examining the model fit of a latent factor model in which diverse desires, diverse beliefs, knowledge, visual perspective taking, false belief, and false sign items load onto a single latent factor. In order to assess how false sign scaled in relation to mental-state constructs, we included false sign in IRT models. Thus, the latent factor estimated includes variance from the five mental-state categories (i.e., diverse desires, diverse beliefs, knowledge, false belief and visual perspective taking) and false sign. Acceptable fit was determined by a χ^2:*df* ratio ≤ 2 (Schreiber et al., 2006), approximate fit indices: Tucker-Lewis index (TLI) and comparative fit index (CFI) $\geq .90$ and comparative fit index (CFI) $\geq .90$ (Brown, 2006), goodness of fit indices (GFI/AGFI) $> .90$, and absolute fit indices: root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR) $\leq .08$ (Hu & Bentler, 1999). Our novel measure of ToM was designed such that stories were locally independent – scores on each story were independent of participants' scores on different stories, and therefore shared variance across stories relates to underlying cognitive abilities (e.g., false sign understanding, ToM, story comprehension, etc.).

We utilized our IRT models to examine item *difficulty*–how items scale from most easy to most difficult, and also assessed item *discrimination*–or how well individual questions within our measure assessed ToM to ensure that items were accurate assessments of participant's underlying ability. Models estimated participants' latent ability (e.g., a standardized latent measurement of participants' performance across mental-state and false sign items), difficulty (e.g., the level of latent ability needed to reach a .5 probability of passing each item, which can be used to order item difficulty across the measure as a whole), and item discrimination (e.g., how well each item assesses ability, or discriminates between participants with different latent abilities, and also how closely an item is related to the underlying latent factor).

We examined scaling both using only prediction items (following work by Wellman and Liu, 2004) and then also for both prediction and explanation questions, given that that the addition of explanation questions was designed to better capture variability in children's ToM.

Scaling of Prediction Items

Data met assumptions to fit a 2 parameter logistic model: data fit a unidimensional factor, and were locally independent because each ordinal item was taken from a single story. Unidimensionality was determined by scree plots of principal components and factor analysis, followed by fitting a single-factor model to the data. Model fit indices were good, χ^2:df = 0.78, TLI = 0.99, CFI = 0.99, RMSEA = 0.02, 95%CI [0.00, 0.03], supporting the data met assumptions for unidimensionality. As described in Scaling of Mental States, items were designed to be locally independent. Thus, data met model assumptions.

The 2 parameter logistic model estimates the probability of children scoring either 0 or 1, points on each item based on the child's underlying latent trait ability, the question difficulty, and the question discrimination.

Although there are no established standards for acceptable discrimination, good discrimination parameters are generally considered to be between 0.5 and 2.5 (Reeves & Fayers, 2005). Items were generally discriminating, in which a ranged from 0.47 to 5.24, supporting the inclusion of these items in our further analyses of the measures as a whole.

Item difficulties are presented in rank-order from easiest to most challenging in Table 2, and ranged from -3.04 to 5.05, however apart from the most challenging item (Visual perspective taking item 6: Cookie), difficulties ranged from -3.04 to 0.58. Thus, items were generally easy, as shown by negative difficulty (b) coefficients (see Table 2), and suggesting that a below-average latent ability was needed to pass prediction items in our sample.

Table 2

Item Difficulties and Discrimination of Prediction Questions

Category	Item description	Difficulty b	Disc. a
Knowledge	Expertise item 2: Cooking	-3.04	1.10
Diverse desires	Item 5: Ball	-2.81	1.05
Visual perspective	Item 3: Giraffe	-2.34	0.58
Visual perspective	Item 4: Raisin box	-2.28	1.03
Diverse desires	Item 2: Picnic	-1.93	1.47
Diverse desires	Item 6: Grapes	-1.88	1.39
Diverse desires	Item 4: Book	-1.85	1.20
Diverse desires	Item 3: Puzzle	-1.82	1.44
Knowledge	Expertise item 3: Car	-1.79	1.88
Diverse desires	Item 1: Hat	-1.78	1.68
Visual perspective	Item 1: Mountain	-1.75	1.98
Knowledge	Expertise item 1: Plane	-1.72	1.05
False sign	Item 3: Airport	-1.68	0.95
Diverse belief	Item 1: Snack	-1.68	1.34
False sign	Item 5: Horse	-1.60	1.79
Diverse belief	Item 4: Cat	-1.45	1.68
False sign	Item 1: Bakery	-1.34	1.56

False sign	Item 6: Ice cream	-1.27	2.03
Visual perspective	Item 2: Bird	-1.24	1.21
Knowledge	Access item 1: Alligator	-1.23	2.39
Knowledge	Access item 2: Spoon	-1.21	2.82
Knowledge	Access item 3: Flower	-1.19	2.40
Diverse belief	Item 3: Gift	-1.19	1.49
False belief	Contents item 3: Cheerios	-1.19	5.24
False belief	Location item 2: Cookie	-1.13	1.72
False sign	Item 2: Carrot	-1.12	1.83
False belief	Contents item 2: Crayon box	-1.06	2.30
Diverse belief	Item 2: Bunny	-1.00	2.23
False belief	Location item 3: Bag	-1.00	1.65
False belief	Contents item 1: Band aid	-0.94	3.45
False sign	Item 4: Cupboard	-0.56	1.05
False belief	Location item 1: Peach	-0.54	1.38
Visual perspective	Item 5: Apple	0.58	1.42
Visual perspective	Item 6: Cookie	5.05	0.47

Note. b = difficulty, a = discrimination, Access = knowledge access, Expert = knowledge expertise, Location = false belief location change, Contents = false belief unexpected contents.

At the broadest level, our results replicate foundational work by Wellman & Liu (2004) and others since (Osterhaus et al., 2022; Peterson et al., 2012; Shahaeian et al., 2011) showing the relatively easier reasoning about desires, diverse beliefs, and knowledge, relative to more difficult false beliefs items. Also following Wellman and Liu (2004), diverse desires items were generally easier compared to diverse beliefs, knowledge, and false beliefs. Our repeated measures of each mental state, and additional mental states not tested in the Wellman & Liu scale also revealed new patterns that expand the scale tested in Wellman and Liu (2004) (see also Peterson et al., 2012). Specifically, diverse beliefs was not consistently easier than knowledge. Reasoning about expertise (not tested in Wellman and Liu, 2004) was easier than diverse beliefs and occurred along-side diverse desires. Knowledge access (included in Wellman and Liu, 2004)

was both easier and harder than diverse belief items, and therefore showed intermediate difficulty. Visual perspective taking (not tested in Wellman and Liu, 2004) showed the largest range of difficulty; it was both among the easiest and hardest tasks, with intermediate difficulty as well. Finally, false sign (also not tested in Wellman and Liu, 2004) was harder than diverse desires, and paralleled difficulty levels of diverse belief and false belief.

Scaling of Prediction and Explanation Items

To examine discrimination and scaling, we fit a graded response model (Samejima, 1969). Our data met assumptions to fit this model in that data were locally independent and unidimensional, as suggested by the scree plot and good model fit of a single-factor model, χ^2:*df* = 0.81, TLI = 0.98, CFI = 0.99, RMSEA = 0.04, 95%CI [0.03, 0.04].

The graded response model estimates the probability of children scoring 0, 1, or 2 points on each item based on the child's underlying latent trait ability, the question difficulty, and the question discrimination. We report results from the graded response model in Table 1, and plot item parameters in Figure 1. We report item difficulty for a score of 0 compared to a score of either 1 or 2 (b1), where difficulty parameters describe how much of the standardized latent trait is needed to earn a score of 1 or 2. We also report item difficulty for a score of 0 and 1 compared to a score of 2 (b2), where difficulty describes how much of the standardized latent trait is needed to earn a score of 2. In addition to reporting the item difficulties extracted from the graded response model, which gives difficulty in relation to all 3 response options (i.e., scores of 0, 1 and 2), we also report generalized item difficulty as described by Ali and colleagues (2015). This generalized difficulty gives a single difficulty parameter for each item, which makes comparisons across items in order to estimate overall difficulty and scaling simpler.

Across all items, discriminations were high, and ranged from 1.07 to 2.35, suggesting that items were accurate discriminators of different latent abilities, and broadly were good measures of ToM and false sign. Item discriminations in this model in which items reflected participants' performance on both prediction and difficulty had higher discrimination than only prediction items alone.

General item difficulties (b) ranged from -1.79 to 2.39, and most difficulty coefficients (b) were negative, similar to findings with prediction items only which also were negative. Unlike the model with prediction items only, the range of difficulties was overall *less* negative, suggesting that adding explanation questions was effective in expanding the ability that can be tested with our measure. Similar to the difficulty parameters extracted from the model of prediction items only, the easiest items to pass were assessments of diverse desires. Also similar to the model of prediction items only, several mental states did not show a clear rank-order in relation to each other: Visual perspective taking, knowledge, diverse beliefs, false belief, and false sign items were harder than diverse desires, but items varied in difficulty. In adding explanation questions, knowledge expertise was no longer easier than knowledge access, suggesting that explanations for knowledge expertise were more challenging than prediction questions alone. In the prediction-questions-only analysis, false belief items were the most challenging, however when adding in explanation questions, false belief items were similar in difficulty to mental states such as visual perspective taking, knowledge and diverse beliefs. This suggests that when adding explanation questions, false belief was overall easier for participants than prediction questions alone. We also looked at variability in task format, and found that reasoning about a character whose belief was discrepant from the participant's own (diverse beliefs about the self versus other) was easier than reasoning about two characters who had

discrepant beliefs (diverse beliefs about a boy and a girl). However, we did not see this pattern in diverse desires, which may be because diverse desires items were generally easy overall.

Table 3

Item Parameters for ToM and False Sign Tasks from the Graded Response Model

				Difficulty			Disc.
Category	Item description	*M*	*SD*	b	b1	b2	a
Diverse desires	Item 2: Picnic	1.75	0.60	-1.79	-2.12	-1.48	1.71
Diverse desires	Item 5: Ball	1.66	0.59	-1.79	-2.71	-0.95	1.28
Diverse desires	Item 3: Puzzle	1.71	0.63	-1.55	-1.97	-1.17	1.74
Diverse desires	Item 1: Hat	1.65	0.64	-1.47	-2.07	-0.87	1.60
False sign	Item 3: Airport	1.57	0.70	-1.36	-1.99	-0.73	1.33
Diverse desires	Item 6: Grapes	1.59	0.69	-1.33	-1.89	-0.75	1.47
Diverse desires	Item 4: Book	1.57	0.71	-1.32	-1.88	-0.74	1.36
False belief	Location item 2: Cookie	1.55	0.69	-1.10	-1.72	-0.46	1.77
Visual perspective	Item 1: Mountain	1.55	0.66	-1.08	-1.78	-0.30	2.00
Diverse beliefs	Item 1: Snack	1.41	0.72	-1.03	-1.88	-0.11	1.39
Knowledge	Expert item 2: Cooking	1.37	0.59	-1.03	-2.40	0.40	1.81
Visual perspective	Item 3: Giraffe	1.41	0.74	-1.03	-1.82	-0.21	1.24
Knowledge	Access item 3: Spoon	1.56	0.75	-0.97	-1.30	-0.60	2.18
False belief	Location item 3: Bag	1.47	0.76	-0.93	-1.45	-0.38	1.57
Knowledge	Access item 2: Flower	1.5	0.77	-0.89	-1.26	-0.46	2.15
Visual perspective	Item 2: Bird	1.44	0.73	-0.84	-1.47	-0.14	1.88
Visual perspective	Item 4: Raisin box	1.33	0.69	-0.84	-1.92	0.30	1.37
False sign	Item 6: Ice cream	1.48	0.74	-0.83	-1.36	-0.22	2.16
False belief	Contents item 1: Band aid	1.48	0.74	-0.81	-1.32	-0.24	2.35
False belief	Contents item 2: Crayon	1.49	0.77	-0.81	-1.19	-0.36	2.26
False sign	Item 2: Carrot	1.4	0.79	-0.75	-1.21	-0.21	1.84
Diverse beliefs	Item 2: Bunny	1.38	0.78	-0.72	-1.28	-0.08	1.65
False sign	Item 5: Horse	1.35	0.68	-0.72	-1.65	0.30	1.92
False belief	Contents item 3: Cheerios	1.38	0.73	-0.71	-1.47	0.12	1.62
Knowledge	Expert item 3: Car	1.31	0.66	-0.70	-1.76	0.47	1.88
False sign	Item 1: Bakery	1.33	0.73	-0.67	-1.46	0.21	1.82
False belief	Location item 1: Peach	1.36	0.76	-0.64	-1.26	0.06	1.91
Diverse beliefs	Item 4: Cat	1.3	0.71	-0.58	-1.43	0.36	1.88
Knowledge	Expert item 1: Plane	1.21	0.74	-0.56	-1.49	0.47	1.32
Diverse beliefs	Item 3: Gift	1.27	0.8	-0.54	-1.15	0.18	1.41
Knowledge	Access item 1: Alligator	1.26	0.72	-0.50	-1.33	0.44	1.93
False sign	Item 4: Cupboard	0.99	0.79	0.03	-0.71	0.93	1.37
Visual perspective	Item 5: Apple	0.83	0.82	0.37	-0.10	1.04	1.69
Visual perspective	Item 6: Cookie	0.31	0.5	2.39	1.13	4.16	1.08

Note. b = general difficulty, as described by Ali and colleagues (2015), b1 = difficulty threshold for 0 vs. 1-2, b2 = difficulty threshold for 0-1 vs. 2, a = discrimination, Access = knowledge

access, Expert = knowledge expertise, Location = false belief location change, Contents = false belief unexpected contents.

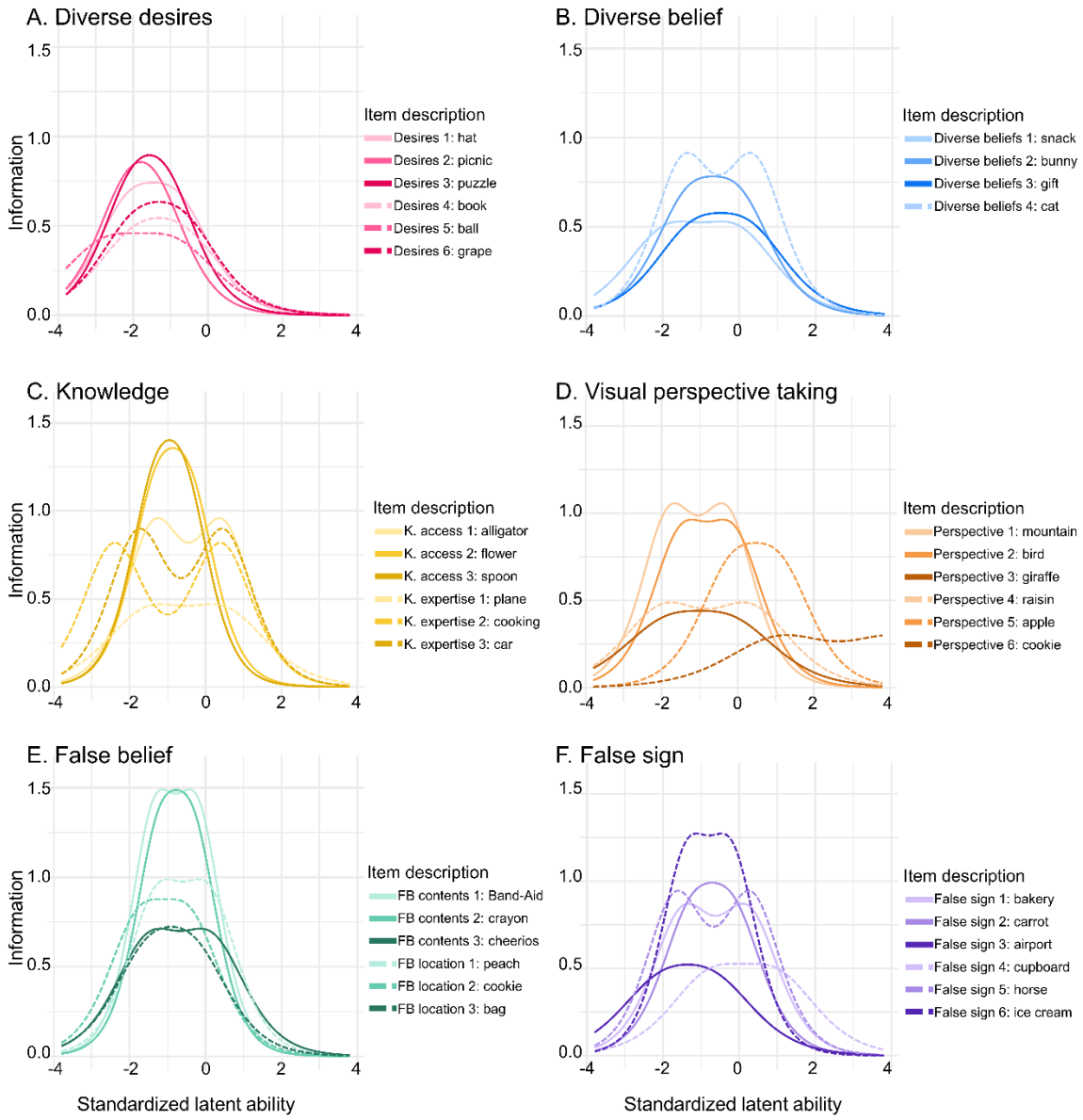
Next, we examined both item information curves (an assessment of the information captured by each item) and the test information function (TIF, an assessment of the information captured by the measures as a whole). Item information curves illustrate which items are most informative psychometrically, meaning that those items contribute most strongly to the estimate of participants' latent ability. Items contain more information if they have a higher discrimination parameter, and therefore items with high discrimination have a higher peak in their information curve (Embretson & Reise, 2000). The TIF shows the test precision at different levels of the latent trait, and is a function of the summed item information curves (Embretson & Reise, 2000). Thus, the TIF shows which latent trait abilities the measure is most accurate at assessing. We also use the TIF in order to identify characteristics of participants for whom the measure was most accurate at assessing.

Item information curves mirrored results from item characteristic curves and the difficulty and discrimination summaries in which diverse desires items were all informative within the same range (around $\theta = -2$), see Figure 3. Similarly, false belief items were also informative within a narrow range of latent ability ($\theta = -1$). Other mental states – diverse beliefs, knowledge and visual perspective taking – and false sign showed that items covered a broader range of abilities, suggesting that there were both easy and difficult items within the category (and also as shown in the difficulty estimates in Table 3). Some items showed bimodal peaks, in which items were informative both at easier difficulties (e.g., if participants earned a 0, the item was informative in estimating their θ) and at higher difficulties (e.g., if participants earned a 2,

the item was informative in estimating their θ). Some mental states – namely diverse desires, diverse beliefs, and false belief – show overlapping information curves across all items, suggesting that items are psychometrically informative within the same range of ability. This may be expected for diverse desires and false belief given that these two mental states showed more cohesive scaling (e.g., diverse desires items were the easiest and also clustered together when examining rank-order of difficulties). Other categories – visual perspective taking, knowledge, and false sign – showed that different items were informative at different levels of ability, which was also reflected in the range of difficulties for items within each of these categories. For example, the least difficult false sign item was the 5th easiest, and the most difficulty item was the 32nd easiest, suggesting that these items would be informative in estimating θ at different abilities.

Figure 3

Item Information Curves for ToM and False Sign



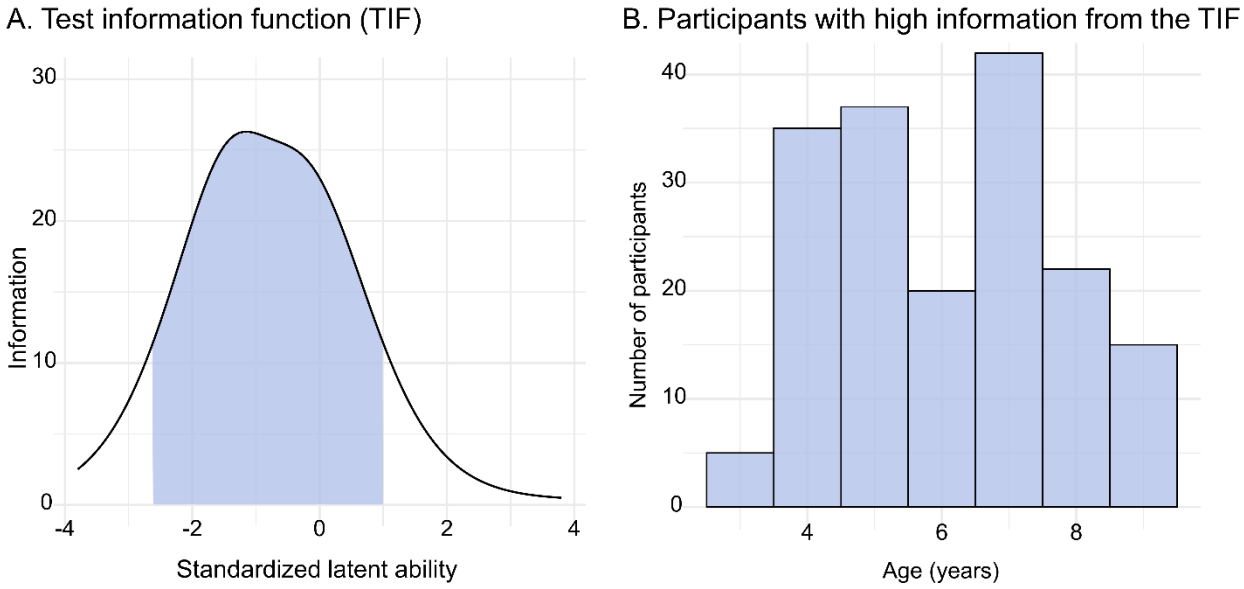
Note. Desires = diverse desires, K. access = knowledge access, K. expertise = knowledge expertise, Perspective = visual perspective taking, FB = false belief.

The TIF showed that our novel comprehensive ToM measure was most informative for participants with a latent trait ability (θ) between -2.5 and 1, which was expected given that most

items had a negative item difficulty, see Figure 3 Panel A. In order to better understand the age range for which this novel measure is most useful, we examined participant characteristics for participants whose latent trait level was between -2.5 and 1. These participants (n = 176) were on average slightly younger than the full sample, M_{age} = 5.94, SD = 1.59, whereas the full sample M_{age} = 6.28, SD = 1.71. However, this subset of participants spanned the full age range of the studied sample, in which the youngest included was 3.1-years-old, and the oldest was 8.8-years-old. These results suggest that this novel measure was able to accurately assess ToM within the full age range tested (3- to 8-years-old).

Figure 4

Test Information Function and Participants with High Information



Note. Panel A illustrates the test information function with -2.5≤θ≤1, showing high information, highlighted. Panel B plots the age of participants with -2.5≤θ≤1.

Discussion

The present study introduces and evaluates a novel measure of children's ToM—the Comprehensive Assessment of Theory of Mind (CAT) and presents novel evidence in how mental states unfold across development. We took a novel approach in administering multiple assessments of each mental state in a large sample, and validated this new measure by examining both reliability (e.g., Cronbach's α, TIF) and validity (e.g., relational validity through correlations with age, vocabulary and executive function). The CAT was accurate in assessing ToM in the tested age range, from 3- to 8-years-old, and is one of only a few multi-mental state measures developed since Wellman and Liu (2004). It also presents a unique approach in its inclusion of a control question, a prediction question, and an explanation question for each item, and in its multiple assessments of each mental state (ranging from 3-6 items for each mental-state category). Importantly, this measure can be administered remotely or in-person, expanding the potential for its utility in the field.

We examined this measure by evaluating both prediction questions alone, and both prediction and explanation questions together as a composite. In general, we replicated the developmental trajectory in ToM described by Wellman and Liu (2004) showing that reasoning about diverse desires was easier than reasoning about knowledge and beliefs. We also expand on prior findings by assessing not only diverse desire, diverse beliefs, knowledge access, and false belief but also knowledge expertise, visual perspective taking, and false sign.

The inclusion of these additional concepts has implications for understanding multiple aspects of ToM. For example, our inclusion of knowledge expertise (e.g., an understanding that different people have different amounts of knowledge about the same topic) revealed that children could more easily predict who had greater expertise than they could predict that a character could lack knowledge that the child themselves have access to (e.g., knowledge

access). Additionally, our findings showed that unlike other mental states, visual perspective taking did not follow a consistent rank-order, with some prediction items being easier and some being among the most challenging (indeed the two most difficult prediction items in our measure assessed visual perspective taking). In addition, we examined how false sign scaled in relation to mental states. False sign developed alongside diverse beliefs and false belief, supporting theoretical (Carey, 1985; Gopnik & Wellman, 1992; 1994) and empirical research (Iao et al., 2011; Leekam et al., 2008; Sabbagh et al., 2006) that points to the importance of representational reasoning as a developmental milestone, and relations between representational reasoning in social and non-social scenarios.

Our examination of both prediction and explanation questions showed that explanations were important in increasing discrimination of questions (discrimination parameters were higher for the graded response model compared to the 2-parameter logistic model with only prediction questions), supporting that explanation questions capture meaningful understanding and variability in children's reasoning. In addition, explanation questions increased the ToM ability that could be assessed with this scale, as apparent in the second 'hump' in some of the item information curves in the graded response model for diverse beliefs, knowledge, visual perspective taking, and false sign. The addition of explanation questions in our measure also revealed new insights into the scaling of mental states: Visual perspective taking has previously been unexplored alongside the traditional set of tasks administered in the ToM measure developed by Wellman and Liu (2004), and visual perspective taking items had a wide range of difficulty, in which some items were among the easiest (after diverse desires), and some were the most difficult. Similarly, knowledge, diverse beliefs, false belief, and false sign items varied in difficulty and did not show a consistent rank-order. In contrast to results with prediction items

only, when including explanation items, false belief items were no longer the most challenging. This finding may have emerged because some children in our sample (e.g., 4- and 5-year-olds) were on the cusp of understanding false belief, and seeing the outcome of the character acting in accordance with their mental state (e.g., going to the oven to get the cookie) may have scaffolded participants' understanding of beliefs that are discrepant from reality. Overall, our data show that the inclusion of explanations captured greater variability in children's mental state and false sign understanding, and increased the discrimination of items. The addition of explanations to each item also raises opportunities to explore how differences in types of explanation questions (e.g., a character performs an action in accordance with their mental state, versus a character says something that reflects their mental state).

Finally, no study has examined difficulty scaling of different item contexts (e.g., whether participants are asked to compare their own mental state versus another, or compare mental states across two different characters), or different aspects of mental-state concepts (e.g., understanding access to knowledge as well as knowledge and expertise). In our novel assessment, the addition of knowledge expertise items was revealing in showing that knowledge expertise predictions were easier than knowledge access predictions. In addition, some mental state assessments asked participants to reason about different mental states between themselves and a character, and some asked participants to reason about mental states that differed between two characters (e.g., between a boy and a girl). Specifically, in evaluating both prediction and explanation items, diverse beliefs items in which the participant reasoned about their own thoughts versus a character's thoughts were easier than items in which they reasoned about two characters with different thoughts. Our multiple assessments of each mental state, inclusion of mental states previously not assessed with the Wellman and Liu measure (2004), inclusion of

both prediction and explanation questions, and our variations in task format led to new insights into ToM development and opened up additional questions for the field regarding the consistency of the ToM scale when there are repeated measures and questions regarding how task structure (e.g., self versus other, two characters with opposing beliefs) may influence scaling. We encourage future use of this measure and also extensions of it exploring additional types of explanations and the addition of mental states (e.g., assessing emotion).

In sum, we present a large-scale study of over 200 children in a wide age range (from 3- to 8-years-old) investigating multiple mental states: diverse desires, diverse beliefs, knowledge access and expertise, false belief, and visual perspective taking. In addition, we assessed false sign understanding. At the broadest level, we replicate prior scaling efforts (Osterhaus et al., 2022; Peterson et al., 2005; Wellman & Liu, 2004) showing that children can predict a character's diverse desires before they can predict that a character will have a false belief. We also expand on this foundational scaling through the inclusion of explanation questions, which illuminated a richer and much more overlapping pattern of difficulty for knowledge, visual perspective taking, diverse beliefs, and false belief. Future research should continue to investigate how question format (e.g., prediction, explanation) and the question content (e.g., explaining a character's actions, explaining a character's verbal response) may illuminate nuances in the difficulty of different mental states. We encourage future use of our multi-assessment scale of multiple mental states, and future expansions of this measure.

References

Ali, U. S., Chang, H., & Anderson, C. J. (2015). Location indices for ordinal polytomous items based on item response theory. *ETS Research Report, 2,* 1-13. https://doi.org/10.1002/ets2.12065

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52,* 388-407. https://doi.org/10.3758/s13428-019-01237-x

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology, 65*(5), 825–839. https://doi.org/10.1080/17470218.2012.676055

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37-46. https://doi.org/10.1016/0010-0277(85)90022-8

Beaudoin, C., Leblanc, E., Gagner, C., & Beauchamp, M. H. (2020). Systematic review and inventory of theory of mind measures for young children. *Frontiers in Psychology, 10,* 2905. https://doi.org/10.3389/fpsyg.2019.02905

Blijd-Hoogewys, E. M. A., Van Geert, P. L. C., Serra, M., & Minderaa, R. B. (2008). Measuring theory of mind in children. Psychometric properties of the ToM storybooks. *Journal of Autism and Developmental Disorders, 38,* 1907-1930. https://doi.org/10.1007/s10803-008-0585-3

Bowman, L. C., Liu, D., Meltzoff, A. N., & Wellman, H. M. (2012). Neural correlates of belief- and desire-reasoning in 7-and 8-year-old children: An event-related potential study. *Developmental Science, 15*(5), 618-632. https://doi.org/10.1111/j.1467-7687.2012.01158.x

Carey, S. (1985). *Conceptual change in childhood*. MIT press.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72*(4), 1032-1053. https://doi.org/10.1111/1467-8624.00333

Carlson, S. M., Mandell, D. J., & Williams, L. (2004). Executive function and theory of mind: stability and prediction from ages 2 to 3. *Developmental Psychology, 40*(6), 1105–1122. https://doi.org/10.1037/0012-1649.40.6.1105

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child Development, 85*(5), 1777–1794. https://doi.org/10.1111/cdev.12237

Fabricius, W. V., & Khalil, S. L. (2003). False beliefs or false positives? Limits on children's understanding of mental representation. *Journal of Cognition and Development, 4*(3), 239-262. https://doi.org/10.1207/S15327647JCD0403_01

Fabricius, W. V., Boyer, T. W., Weimer, A. A., & Carroll, K. (2010). True or false: Do 5-year-olds understand belief? *Developmental Psychology, 46*(6), 1402–1416. https://doi.org/10.1037/a0017648

Flavell, J. H., Shipstead, S. G., & Croft, K. (1978). Young children's knowledge about visual perception: Hiding objects from others. *Child Development, 49*(4), 1208-1211. https://doi.org/10.2307/1128761

Ford, R. M., Driscoll, T., Shum, D., & Macaulay, C. E. (2012). Executive and theory-of-mind contributions to event-based prospective memory in children: Exploring the self-

projection hypothesis. *Journal of Experimental Child Psychology, 111*(3), 468-489. https://doi.org/10.1016/j.jecp.2011.10.006

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*(4), 531-534.

Gerrans, P., & Stone, V. E. (2008). Generous or parsimonious cognitive architecture? Cognitive neuroscience and theory of mind. *The British Journal for the Philosophy of Science, 59*(2), 121-141. https://doi.org/10.1093/bjps/axm038

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language, 7*(1-2), 145-171.

Gopnik, A., & Wellman, H. M. (1994). 10 the theory theory. *Mapping the mind:*

Gopnik, A., Slaughter, V., & Meltzoff, A. (1994). Changing your views: How understanding visual perception can lead to a new theory of the mind. In *Children's early understanding of mind: Origins and development*, 157-181.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Iao, L. S., Leekam, S., Perner, J., & McConachie, H. (2011). Further evidence for nonspecificity of theory of mind in preschoolers: Training and transferability in the understanding of false beliefs and false signs. *Journal of Cognition and Development, 12*(1), 56-79. https://doi.org/10.1080/15248372.2011.539523

Kaufman, A. S., & Kaufman, N.L. (2004). Kaufman brief intelligence test–second edition (KBIT-2). *Circle Pines, MN: American Guidance Service*.

Kramer, H. J., Lagattuta, K. H., & Sayfan, L. (2015). Why is happy–sad more difficult? Focal emotional information impairs inhibitory control in children and adults. *Emotion, 15*(1), 61–72. https://doi.org/10.1037/emo0000023

Lane, J. D., Wellman, H. M., & Evans, E. M. (2014). Approaching an understanding of omniscience from the preschool years to early adulthood. *Developmental Psychology, 50*(10), 2380–2392. https://doi.org/10.1037/a0037715

Leekam, S., Perner, J., Healey, L., & Sewell, C. (2008). False signs and the non-specificity of theory of mind: Evidence that preschoolers have general difficulties in understanding representations. *British Journal of Developmental Psychology, 26*(4), 485-497. https://doi.org/10.1348/026151007X260154

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of mind development in Chinese children: A meta-analysis of false-belief understanding across cultures and languages. *Developmental Psychology, 44*(2), 523–531. https://doi.org/10.1037/0012-1649.44.2.523

Liu, D., Sabbagh, M. A., Gehring, W. J., & Wellman, H. M. (2009). Neural correlates of children's theory of mind development. *Child Development, 80*(2), 318-326. https://doi.org/10.1111/j.1467-8624.2009.01262.x

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development, 78*(2), 622–646. https://doi.org/10.1111/j.1467-8624.2007.01018.x

Muris, P., Steerneman, P., Meesters, C., Merckelbach, H., Horselenberg, R., van den Hogen, T., & van Dongen, L. (1999). The TOM test: A new instrument for assessing theory of mind in normal children and children with pervasive developmental disorders. *Journal of*

Autism and Developmental Disorders, 29, 67-80. https://doi.org/10.1023/A:1025922717020

Osterhaus, C., Koerber, S. and Sodian, B. (2016), Scaling of advanced theory-of-mind tasks. *Child Development, 87*(6), 1971-1991. https://doi.org/10.1111/cdev.12566

Osterhaus, C., Kristen-Antonow, S., Kloo, D., & Sodian, B. (2022). Advanced scaling and modeling of children's theory of mind competencies: Longitudinal findings in 4- to 6-year-olds. International Journal of Behavioral Development, 46(3), 251-259. https://doi.org/10.1177/01650254221077334

Parkin, L. J. (1994). *Children's understanding of misrepresentation* (Doctoral dissertation, University of Sussex).

Perner, J., Frith, U., Leslie, A. M., & Leekam, S. R. (1989). Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child Development*, 689-700. https://doi.org/10.2307/1130734

Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Development, 83*(2), 469-485. https://doi.org/10.1111/j.1467-8624.2011.01728.x

Piaget, J., & Inhedler, B. (1969). *The psychology of the child.* Basic Books.

Pillow, B. H. (1989). Early understanding of perception as a source of knowledge. *Journal of Experimental Child Psychology, 47*(1), 116-129. https://doi.org/10.1016/0022-0965(89)90066-0

Pratt, C., & Bryant, P. (1990). Young children understand that looking leads to knowing (so long as they are looking into a single barrel). *Child Development, 61*(4), 973-982. https://doi.org/10.2307/1130869

R Core Team. (2019). R (Version 3.6.2.). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. In P.M. Fayers & R. D. Hays (Eds.). *Assessing Quality of Life in Clinical Trials: Methods and Practice.* Oxford University Press.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental Psychology*, *33*(1), 12. https://doi.org/10.1037//0012-1649.33.1.12

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications, 9*(1), 1027. https://doi.org/10.1038/s41467-018-03399-2

Rizopoulos, D. (2006). ltm: an R package for latent variable modeling and item response analysis. *Journal of Statistical Software, 17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Ross-Sheehy, S., Schneegns, S., & Spencer, J. P. (2015). The infant orienting with attention task: Assessing the neural basis of spatial attention in infancy. *Infancy*, *20*(5), 467-506. https://doi.org/10.1111/infa.12087

Sabbagh, M. A., Moses, L. J., & Shiverick, S. (2006). Executive functioning and preschoolers' understanding of false beliefs, false photographs, and false signs. *Child Development*, *77*(4), 1034-1049. https://doi.org/10.1111/j.1467-8624.2006.00917.x

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, 17.* Richmond, VA: Psychometric Society.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*(6), 323-338. https://doi.org/10.3200/JOER.99.6.323-338

Shahaeian, A., Peterson, C. C., Slaughter, V., & Wellman, H. M. (2011). Culture and the sequence of steps in theory of mind development. *Developmental Psychology, 47*(5), 1239–1247. https://doi.org/10.1037/a0023899

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 73-191.

Todd, A. R., Cameron, C. D., & Simpson, A. J. (2017). Dissociating processes underlying level-1 visual perspective taking in adults. *Cognition, 159*, 97-101. https://doi.org/10.1016/j.cognition.2016.11.010

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*(2), 523-541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655-684. https://doi.org/10.1111/1467-8624.00304

Wellman, H. M., Fang, F., Liu, D., Zhu, L., & Liu, G. (2006). Scaling of theory-of-mind understandings in Chinese children. *Psychological Science, 17*(12), 1075-1081. https://doi.org/10.1111/j.1467-9280.2006.01830.x

Wellman, H. M., Fang, F., & Peterson, C. C. (2011). Sequential progressions in a theory-of-mind scale: Longitudinal perspectives. *Child Development, 82*(3), 780-792. https://doi.org/10.1111/j.1467-8624.2011.01583.x

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103-128. https://doi.org/10.1016/0010-0277(83)90004-5

Appendices

Table S1

Children's Item-level ToM and False Sign Performance across Development

		3-year-olds		4-year-olds		5-year-olds		6-year-olds		7-year-olds		8-year-olds	
		N	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*
Diverse desires	Item 1	23	0.87 (0.81)	31	1.42 (0.76)	34	1.82 (0.46)	35	1.83 (0.45)	34	1.79 (0.54)	39	1.87 (0.34)
	Item 2	25	1.04 (0.84)	31	1.65 (0.61)	34	1.82 (0.58)	35	1.77 (0.65)	35	2.00 (0.00)	39	2.00 (0.00)
	Item 3	23	1.00 (0.85)	30	1.43 (0.77)	34	1.79 (0.54)	35	1.86 (0.49)	34	1.91 (0.38)	39	1.97 (0.16)
	Item 4	25	0.84 (0.80)	34	1.21 (0.81)	36	1.69 (0.52)	35	1.74 (0.66)	34	1.82 (0.46)	40	1.85 (0.48)
	Item 5	25	1.08 (0.57)	34	1.44 (0.70)	36	1.64 (0.64)	35	1.83 (0.45)	35	1.94 (0.34)	40	1.85 (0.43)
	Item 6	25	0.80 (0.65)	34	1.32 (0.84)	36	1.75 (0.55)	35	1.80 (0.53)	35	1.71 (0.62)	40	1.88 (0.40)
Diverse beliefs	Item 1	21	0.76 (0.62)	24	1.04 (0.81)	18	1.44 (0.70)	23	1.91 (0.29)	21	1.52 (0.60)	31	1.68 (0.60)
	Item 2	25	0.40 (0.65)	31	0.94 (0.81)	35	1.43 (0.78)	35	1.60 (0.60)	35	1.74 (0.44)	37	1.81 (0.46)
	Item 3	25	0.28 (0.54)	31	1.03 (0.84)	35	1.17 (0.79)	35	1.49 (0.66)	34	1.68 (0.59)	39	1.64 (0.58)
	Item 4	25	0.44 (0.58)	34	0.97 (0.72)	35	1.26 (0.56)	35	1.37 (0.60)	35	1.63 (0.55)	40	1.80 (0.46)
	Item 5	26	0.31 (0.47)	33	0.94 (0.75)	36	1.25 (0.55)	35	1.46 (0.61)	34	1.41 (0.56)	40	1.85 (0.36)
	Item 6	25	0.44 (0.71)	34	1.15 (0.82)	35	1.57 (0.70)	35	1.69 (0.68)	35	1.86 (0.36)	40	1.95 (0.22)

Children's Item-level ToM and False Sign Performance across Development cont.

		3-year-olds		4-year-olds		5-year-olds		6-year-olds		7-year-olds		8-year-olds	
		N	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*
Knowledge	Item 1	24	0.42 (0.72)	34	1.35 (0.85)	35	1.69 (0.58)	35	1.80 (0.53)	35	1.80 (0.47)	39	1.92 (0.35)
	Item 2	22	0.64 (0.66)	26	0.77 (0.51)	21	1.33 (0.66)	24	1.29 (0.75)	21	1.48 (0.75)	31	1.65 (0.61)
	Item 3	26	0.88 (0.33)	34	0.91 (0.45)	35	1.40 (0.50)	35	1.43 (0.61)	35	1.63 (0.55)	40	1.77 (0.42)
	Item 4	25	0.60 (0.50)	34	0.85 (0.56)	36	1.31 (0.58)	35	1.54 (0.56)	35	1.66 (0.48)	40	1.62 (0.54)
	Item 5	26	0.69 (0.62)	32	1.19 (0.74)	35	1.46 (0.61)	35	1.89 (0.32)	33	1.88 (0.33)	40	1.90 (0.38)
	Item 6	25	0.64 (0.64)	33	0.94 (0.79)	36	1.47 (0.70)	35	1.69 (0.58)	34	1.74 (0.45)	40	1.88 (0.40)
Visual perspective	Item 1	26	0.73 (0.53)	34	0.82 (0.80)	35	1.49 (0.70)	35	1.77 (0.49)	35	1.69 (0.63)	40	1.75 (0.49)
	Item 2	26	0.42 (0.50)	34	1.12 (0.54)	36	1.28 (0.66)	34	1.47 (0.56)	34	1.62 (0.49)	39	1.79 (0.52)
	Item 3	24	0.08 (0.28)	33	0.21 (0.42)	36	0.56 (0.65)	35	1.11 (0.76)	35	1.31 (0.80)	40	1.38 (0.70)
	Item 4	25	0.08 (0.28)	34	0.18 (0.46)	36	0.19 (0.40)	35	0.37 (0.55)	35	0.34 (0.48)	40	0.60 (0.59)
	Item 5	23	0.87 (0.81)	31	1.42 (0.76)	34	1.82 (0.46)	35	1.83 (0.45)	34	1.79 (0.54)	39	1.87 (0.34)
	Item 6	25	1.04 (0.84)	31	1.65 (0.61)	34	1.82 (0.58)	35	1.77 (0.65)	35	2.00 (0.00)	39	2.00 (0.00)

Children's Item-level ToM and False Sign Performance Across Development cont.

		3-year-olds		4-year-olds		5-year-olds		6-year-olds		7-year-olds		8-year-olds	
		N	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*
False belief	Item 1	26	0.42 (0.58)	34	1.00 (0.85)	36	1.69 (0.52)	35	1.80 (0.41)	34	1.68 (0.53)	40	1.92 (0.35)
	Item 2	26	0.54 (0.76)	34	0.94 (0.89)	36	1.58 (0.69)	35	1.83 (0.38)	35	1.89 (0.32)	40	1.85 (0.43)
	Item 3	25	0.36 (0.64)	34	1.12 (0.84)	36	1.56 (0.61)	35	1.69 (0.47)	35	1.54 (0.51)	40	1.68 (0.47)
	Item 4	26	0.42 (0.64)	34	1.06 (0.81)	36	1.39 (0.69)	35	1.51 (0.56)	33	1.67 (0.60)	40	1.82 (0.45)
	Item 5	26	0.62 (0.75)	34	1.32 (0.81)	36	1.56 (0.65)	35	1.77 (0.43)	35	1.89 (0.32)	40	1.88 (0.33)
	Item 6	24	0.54 (0.59)	33	1.06 (0.86)	36	1.50 (0.70)	35	1.69 (0.63)	35	1.83 (0.51)	40	1.85 (0.43)
False sign	Item 1	26	0.54 (0.58)	34	0.82 (0.63)	36	1.31 (0.71)	35	1.54 (0.61)	34	1.62 (0.55)	40	1.88 (0.33)
	Item 2	25	0.48 (0.59)	34	1.00 (0.89)	36	1.53 (0.70)	34	1.62 (0.70)	34	1.65 (0.65)	40	1.82 (0.45)
	Item 3	25	0.88 (0.83)	33	1.21 (0.74)	36	1.69 (0.62)	35	1.69 (0.58)	35	1.83 (0.45)	39	1.87 (0.47)
	Item 4	26	0.31 (0.47)	33	0.58 (0.61)	36	1.06 (0.71)	35	1.11 (0.80)	35	1.34 (0.73)	40	1.30 (0.79)
	Item 5	25	0.64 (0.57)	34	0.97 (0.67)	36	1.39 (0.60)	35	1.60 (0.65)	34	1.50 (0.56)	40	1.73 (0.45)
	Item 6	24	0.54 (0.66)	34	1.09 (0.79)	36	1.39 (0.73)	35	1.80 (0.53)	35	1.74 (0.51)	40	1.92 (0.27)

Table S2

Children's Average ToM and False Sign Performance across Development

	Age						Possible range
	3-year-olds	4-year-olds	5-year-olds	6-year-olds	7-year-olds	8-year-olds	
N	26	34	36	35	35	40	
Age (years)	3.67 (0.20)	4.54 (0.27)	5.54 (0.19)	6.63 (0.15)	7.60 (0.15)	8.65 (0.11)	3.00–8.99
Vocabulary	13.27 (5.03)	14.29 (4.25)	18.29 (3.48)	21.89 (3.40)	25.88 (4.76)	29.43 (4.87)	0–60
Diverse desires	0.92 (0.53)	1.41 (0.50)	1.75 (0.30)	1.80 (0.24)	1.87 (0.22)	1.90 (0.18)	0–2
Diverse beliefs	0.46 (0.44)	0.98 (0.57)	1.31 (0.44)	1.56 (0.30)	1.67 (0.33)	1.75 (0.40)	0–2
Knowledge	0.54 (0.37)	1.01 (0.44)	1.42 (0.38)	1.55 (0.36)	1.65 (0.30)	1.80 (0.22)	0–2
Visual perspective	0.45 (0.48)	0.74 (0.32)	1.07 (0.32)	1.38 (0.28)	1.41 (0.31)	1.55 (0.29)	0–2
False belief	0.48 (0.41)	1.08 (0.65)	1.55 (0.40)	1.71 (0.23)	1.75 (0.29)	1.83 (0.29)	0–2
False sign	0.56 (0.40)	0.95 (0.48)	1.39 (0.43)	1.56 (0.35)	1.62 (0.39)	1.75 (0.26)	0–2

Notes. Means and standard deviations, *M* (*SD*), are shown for a composite of each of the mental-state categories. Means were calculated as children's average score out of all items they had completed within the category.

Table S3

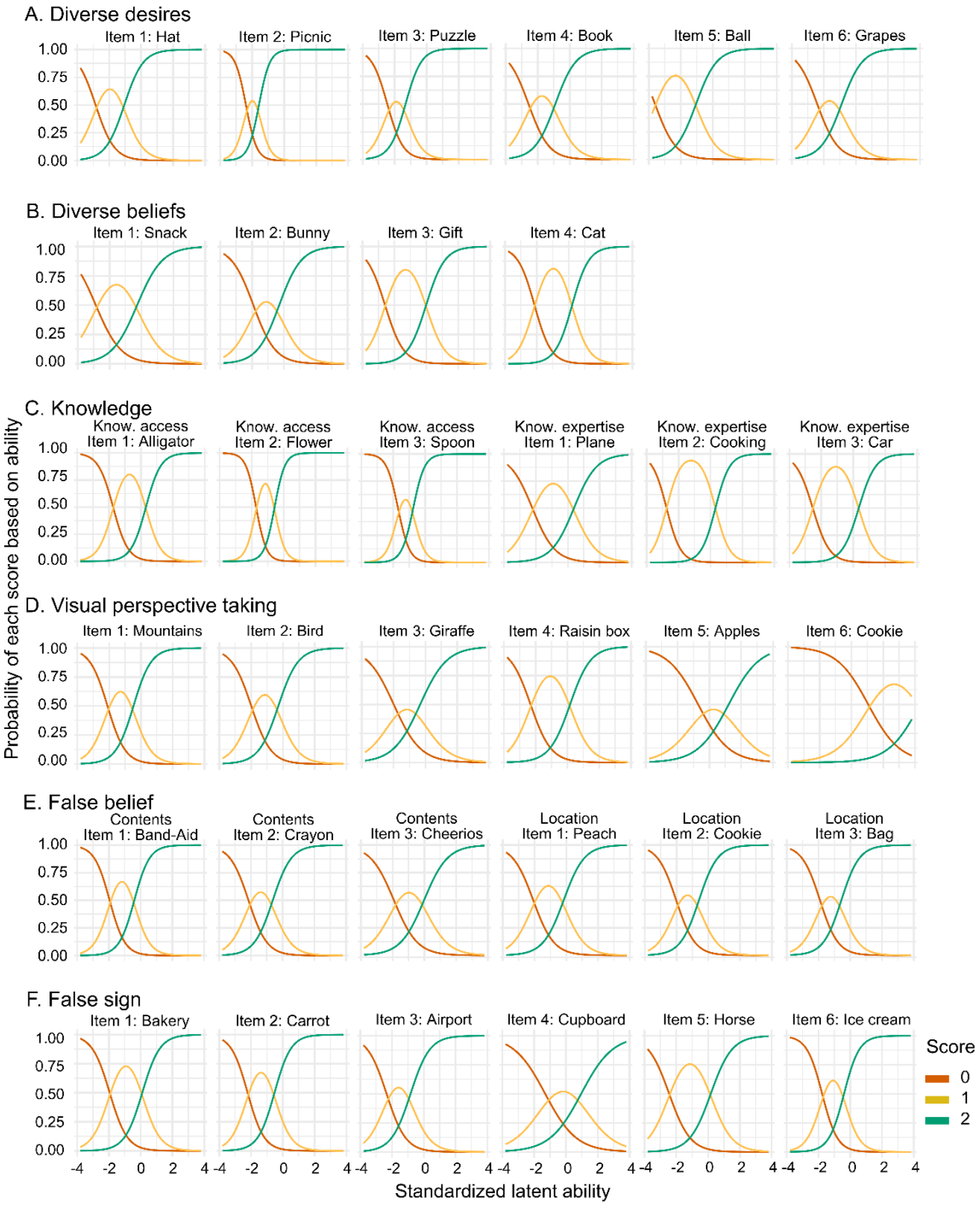
Internal Consistency of ToM Subscales

	N items	α	Inter-item correlation
All ToM items	28	.96	.43
Diverse desires	6	.79	.39
Diverse beliefs	4	.76	.45
Knowledge	6	.87	.54
Visual perspective	6	.77	.36
False belief	6	.88	.56
False sign	6	.83	.46

Notes. N items = number of items within each subscale, α = Cronbach's α, inter-item correlation = mean correlation between all items within each subscale.

Figure S1

Item Characteristic Curves for ToM and False Sign Tasks from the Graded Response Model



Note. Know. access = knowledge access, know. expertise = knowledge expertise. Lines represent the ICC for each score (i.e., 0, 1 and 2 points). The x-axis position of each line reflects the probability of earning that score based on the participant's latent ability, and the slope of each line reflects the discrimination parameter, in which steeper slopes indicate that items were more accurate at measuring the latent ability