Measures of individual differences in adult theory of mind

Yeung, Kit; Apperly, Ian; Devine, R.T.

DOI:
10.1016/j.neubiorev.2023.105481

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Yeung, K, Apperly, I & Devine, RT 2024, 'Measures of individual differences in adult theory of mind: A systematic review', *Neuroscience and biobehavioral reviews*, vol. 157, 105481.
https://doi.org/10.1016/j.neubiorev.2023.105481

Neuroscience and Biobehavioral Reviews 157 (2024) 105481

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Measures of individual differences in adult theory of mind: A systematic review

Elaine Kit Ling Yeung *, Ian A. Apperly, Rory T. Devine

School of Psychology, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom

ARTICLE INFO

Keywords:
Theory of mind
Advanced theory of mind
Measurement
Systematic review
Individual differences
Mentalising

ABSTRACT

Theory of mind (ToM), the ability to understand and reason about mental states, has been extensively studied in young children and clinical populations. A growing interest in examining ToM in adults has emerged over the past two decades, but the extent to which existing measures are suitable for studying adults, especially in detecting individual differences, remains understudied. In this systematic review of 273 studies, 75 measures used to investigate individual differences in adults' ToM were identified. Their sensitivity to individual differences, reliability, and validity were examined. Results suggest that ceiling effects were prevalent, and there was limited evidence to establish the reliability or validity of these measures due to the lack of reports of psychometric properties. Interrelations among measures were inconsistent. These findings highlight the need for future empirical and theoretical work to broaden the evidence base regarding psychometric properties of measures, to develop new measures, and to lay out more specific hypotheses about the relevance of ToM for different social outcomes.

1. Introduction

Theory of mind (ToM), also commonly referred to as mindreading or mentalising, is the ability to represent mental states, reason about them, and make use of them to predict and explain behaviour (Apperly, 2010; Baron-Cohen et al., 1985; Premack and Woodruff, 1978). It is regarded an important ability that facilitates social interaction (e.g., Brüne, 2005; Happe, Frith, 1996; Paal and Bereczkei, 2007; Watson et al., 1999). Early research on the topic focused on ToM development in early childhood (e.g., Gopnik and Astington, 1988; Perner et al., 1987; Wimmer and Perner, 1983) and in people with clinical conditions, especially autism (e.g., Baron-Cohen, 1985; Yirmiya et al., 1998; Hughes et al., 2000). There is now clear evidence that ToM development continues across middle childhood and adolescence (e.g., Apperly et al., 2011; Devine and Hughes, 2013; Hughes, 2016; see Devine, 2021, and Weimer et al., 2021, for a review). Alongside developmental work on children and adolescents, studies of ToM in neurotypical adults, focused on underlying cognitive and neural processes and the presence of individual differences, have also emerged (e.g., Apperly, 2010; Bradford et al., 2015; Mahy et al., 2014; Qureshi et al., 2020; Schurz et al., 2014). Despite this ongoing interest in ToM, there is little consensus on how best to measure individual differences in ToM in neurotypical adults. In this systematic review, we identify two major challenges in measuring individual differences in ToM performance in neurotypical adults, identify existing measures, and critically examine the measurement characteristics of these measures. The over-arching aim of this review is to take stock of work needed to evaluate existing measures and to develop new ones.

1.1. Studying ToM in adults

Research on ToM in adults has proliferated in the previous two decades (Apperly, 2021). Neurotypical adults are considered developmentally mature in their understanding of mental state concepts (Apperly et al., 2009b; Karmakar and Dogra, 2019), providing a baseline for comparison with other populations such as children and clinical groups. However, adults still show patterns of performance on ToM tasks that are analogous to those observed in children, such as demonstrating egocentric biases when they need to take the perspective of a less-informed person (Keysar et al., 2000, 2003), and making inaccurate mental inferences of what another person thinks or feels (Ickes et al., 2000), with notable variation in performance between individuals. From such observations, the study of individual differences in adult ToM performance has emerged as a meaningful research topic. For example,

* Corresponding author.
E-mail address: kxy090@student.bham.ac.uk (E.K.L. Yeung).

https://doi.org/10.1016/j.neubiorev.2023.105481
Received 3 October 2023; Received in revised form 11 November 2023; Accepted 23 November 2023
Available online 28 November 2023

researchers have suggested various sources of such individual differences, including the ability to locate a mind within a mind-space (Conway et al., 2020), or the flexibility to make mental inferences based on varying contexts (Devine, 2021; Hughes and Devine, 2015). There is also research that teases apart adults' ToM ability to make accurate mental inferences and their propensity, or motivation, to use their ToM (Apperly and Wang, 2021; Carpenter et al., 2016; Devine and Apperly, 2022).

Furthermore, researchers have investigated whether adults' ToM performance correlates with various social skills, cognitive abilities, and traits related to psychiatric and neurodevelopmental conditions (e.g., Abu-Akel et al., 2015; German and Hehman, 2006; McGarry et al., 2021; Nilsen and Duong, 2013; Weinstein et al., 2022). Critically, however, the research described above requires that individual differences in ToM in adults can be reliably and validly measured. There are two problems that should raise concerns about current measures.

1.1.1. *Problem 1: measures may not be sensitive to variance in performance in neurotypical adults*

Many studies of individual differences in adults have either employed tasks originally designed for children or for investigating differences between neurotypical adults and adults with psychiatric or neurodevelopmental conditions. According to one account, children acquire an understanding of mental concepts sequentially (Wellman and Liu, 2004). The first concepts include desire, belief and emotion, and subsequent studies suggest that more complex concepts, such as belief-desire reasoning, require the integration of simpler mental state concepts. Empirically, children perform well on all concepts by middle childhood, leaving little possibility of variation in adults. For example, Peterson et al. (2012) found that half of the children aged between 6 and 7.5 passed the hidden emotions task, the most difficult task in the 5-step Theory of Mind Scale (Wellman and Liu, 2004), and 79% children aged between 7.5 and 11.5 were able to pass it. Moreover, the dominant theoretical interpretation considers these findings to chart the acquisition of the concepts that adults are presumed to possess (Peterson et al., 2012; Wellman and Liu, 2004). This interpretation has no capacity to explain variation in the performance of older children and adults, other than as measurement errors in assessing their underlying conceptual competence (Apperly, 2012). If the source of variation in performance on theory-of-mind tasks is indeed measurement error, then individual differences in performance should not be associated with meaningful outcomes (e.g., Hughes and Devine, 2015). However, drawing on research showing that on individual differences in ToM performance in early and middle childhood exhibit rank-order stability over time and correlate with real-world social outcomes such as social competence (e.g., Devine, 2021; Devine et al., 2016; Hughes and Devine, 2015), it is more likely that these individual differences are meaningful, rather than mere measurement errors. Whether the measures used to test older children and adolescents are still sensitive to variation in adults and whether these meaningful individual differences persist into adulthood warrants further research, the above findings indicate that the application of methods that focus on detecting developmental differences to assess individual differences in adults (e.g., El Haj et al., 2017) should be viewed with some caution, as we can expect that they are likely to mask any variation in adults' ToM performance.

An analogous problem exists for tasks designed to detect differences between experimental conditions or between clinical and non-clinical groups. A well-designed task for comparing between different experimental conditions aims to minimise between-participant variation to maximise sensitivity to detect between-condition differences (Hedge, Powell, and Sumner, 2018). By extension, a task designed to be sensitive for detecting differences between clinical and neurotypical populations is also unlikely to be optimised for detecting individual differences within groups. Although tasks designed on this basis may still be good measures of individual differences within neurotypical adults this should not be taken for granted.

1.1.2. *Problem 2: psychometric properties of measures*

Classical test theory provides a framework for evaluating the quality of measures of psychological constructs such as ToM (Fu et al., 2023) and has been applied in to evaluate measures of children's ToM (e.g., Hughes et al., 2000; Devine and Hughes, 2016). According to the classical test theory, a true score on a construct can be approximated by taking repeated measures of it (e.g., Rust et al., 2020). Reliability is characterised by the extent to which repeated measures correlate with one another, as it captures the variance that is not attributed to measurement error of individual tests. Assuming all items in the same measure capture the same construct, the items should correlate with one another, and hence show good internal consistency (Fu et al., 2023; Revelle and Condon, 2019). Even if items present different contexts or settings, or even have different levels of difficulty, internal consistency is expected if the items capture the same underlying construct (e.g. Devine and Hughes, 2016). Internal consistency is often estimated and indicated by standardised reliability coefficients, such as Cronbach's alpha and omega, that can be compared across different studies (Revelle and Condon, 2019). Another type of reliability is test-retest reliability. To the extent that ToM is a trait-like ability (e.g., Devine, 2021), ToM performance should be stable over short periods of time without much fluctuation in rank order and should therefore demonstrate test-retest reliability (Rust et al., 2020). Finally, when task scores are coded from open-ended responses, inter-rater reliability should be examined to ensure the scoring schemes are interpreted and applied in the same way across coders (e.g., Devine, Kovatchev, Grumley Traynor, Smith & Lee, 2023).

A test is considered valid if it measures the construct it is intended to capture. Validity is a matter of degree and is informed by theoretical predictions about how a given construct should behave (Nunnally, 1978). Criterion-related validity concerns how well the measure predicts criterion variables, which are relevant but operationally distinct from the measure itself (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014), making use of the assumption that test performance has practical and theoretical implications. Hence, individual differences in test performance should be associated with behaviour, traits, psychological processes, or performance in other constructs of interest. For example, ToM is assumed to be a keystone social cognitive ability and so should be related to social outcomes (e.g., Banerjee et al., 2011; Canty et al., 2017; Imuta et al., 2016; Devine et al., 2023). Convergent validity makes use of available measures that are viewed as measuring similar or identical constructs to the target construct under consideration. For ToM, this might involve examining associations between measures of ToM that use different stimuli or response formats, and other measures of social cognition. Discriminant validity is supported if the measure captures what is intended to assess, but not other constructs (Rönkkö and Cho, 2022). It is important to examine the discriminant validity of a measure as it establishes what is captured by ruling out what it does not capture.

1.2. *The current study*

Related reviews have focused on early childhood (Beaudoin et al., 2020; Fu et al., 2023; Ziatabar Ahmadi et al., 2015), middle childhood and adolescence, or were limited to literature that assessed alexithymia alongside ToM (Pisani et al., 2021), or did not examine the psychometric properties of measures for adults (Derksen et al., 2018; Osterhaus and Bosacki, 2022). The current study provides the first systematic review and synthesis of measures of ToM that have been adopted to investigate individual differences in neurotypical adults and assesses the appropriateness of measures for use in research on individual differences in ToM performance in adults. We first summarise existing measures that have been adopted to test individual differences in ToM in neurotypical adults. We focus on the age range of 18–65 because ToM processes in older adults beyond 65 can be different from that of younger adults due

to ageing (e.g., Henry, Philips, Ruffman & Bailey, 2013). We analyse the evidence for the reliability and validity of each measure and examine interrelations among these measures. Finally, we discuss the differences between these measures and measures that are used to assess ToM in children.

2. Method

2.1. Search method and selection criteria

A systematic search of relevant empirical papers published between the year 1978 (the year in which Premack and Woodruff first coined the term 'theory of mind') and January 2022 was conducted by accessing the following databases: Scopus, PsycINFO, and Web of Science on 18th January, 2022. The search terms used for searching in Scopus and Web of Science were: ("theory of mind" OR mentali?ing OR "mind reading" OR "mind perception" OR "cognitive empathy" OR "empathic accuracy" OR "mental state attribution" OR "folk psycholog*" OR "perspective taking" OR "false belief*" OR "advanced theory of mind" OR {belief-desire}) AND (adult* OR "beyond childhood" OR "lifespan" OR adolescen*). We conducted the search on PsycINFO using a combination of subject headings and search terms. We searched for entries under the subject headings "theory of mind", "false beliefs", or "mentalization", in addition to those including the search terms (cognitive empathy or empathic accuracy or mind perception). The full search strategy and search timeline can be found in our preregistration on the Open Science Framework (OSF). Our search resulted in 14474 initial results published in English and other languages. After removing duplicates, 9434 papers were retained, out of which 8872 were excluded after a screening of abstracts, due to using only self-report measures, irrelevance (e.g. the search term "false belief*" generated papers referring to fallacious beliefs about the world), a focus on neural activity, absence of neurotypical adult group, or lack of availability in English. Full text of the remaining 562 papers were accessed and checked for eligibility. The final number of reports included in the review was 248, comprising of 273 studies. It was noted that some of the studies adopted more than one measure to be included in the review. The screening process is summarised in the flowchart (Fig. 1) following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement (Page et al., 2021), and the review was preregistered on OSF prior to data analysis.

The current review focuses on behavioural measures of individual differences in ToM in neurotypical adults. Hence, the following inclusion/exclusion criteria were adopted. We included empirical papers that included at least one group of adult participants who did not report any psychiatric or neurophysiological condition, and reported at least one correlation between ToM performance and a behavioural, self-report, or

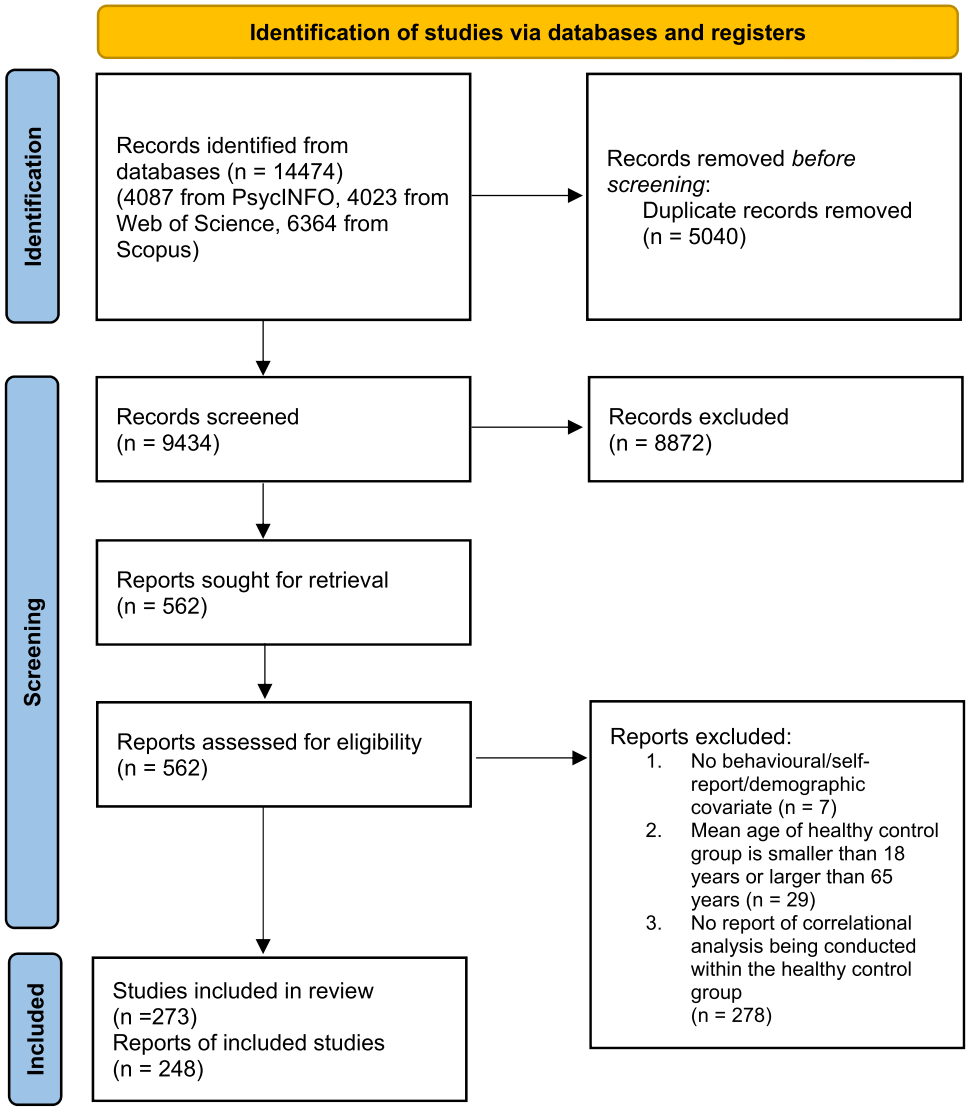

Fig. 1. Flow diagram of study inclusion based on PRISMA (Page et al., 2021).

demographic variable, or else focused on the psychometric properties of the ToM measure(s). Papers that only compared between groups without correlating participants' ToM performance with any other variables were excluded to limit the number of reports included to a manageable number, and due to the reason that these papers did not contribute additional information about the test-retest reliability, convergent validity, and criterion-related validity of the reported measures without running any correlational analyses. Excluding these papers minimised the risk of overshadowing the smaller proportion of reports that showed evidence relevant to test-retest reliability, convergent validity, and criterion-related validity of the measures. Furthermore, we only included papers that included the keywords "theory of mind", "mindreading", "mentalis/zing", or "attribution of mental states" in the current review. Papers that referred to "mentalis/zation" as mind-mindedness or mind-perception were excluded, as these terms refer to the awareness or perception that other human or non-human objects have a mind without necessarily probing into the ability to infer and make use of the information about what is held in the mind of someone. We included studies that measured ToM behaviourally and excluded studies that manipulated ToM between different conditions, or measured ToM in terms of neural activity. Only studies with at least one group of neurotypical participants whose mean age was between 18 and 65 years were included. Studies that examined visual perspective taking were included only if an agent with a perspective different to the participants was presented such that participants had to take the perspective of the agent, to rule out paradigms that only required mental rotation into an alternative spatial position. Studies using only self-perceived measures of ToM were also excluded as meta-analysis results showed minimal correlation between self-reports and behavioural measures of cognitive empathy, a construct commonly defined as a component of theory of mind (Murphy and Lilienfeld, 2019). Reliability of the list of criteria was checked by a second coder screening a subset of 50 papers. The agreement between the two coders was 90%, suggesting good reliability. Discrepancy between screeners were resolved by discussion until mutual agreement was achieved.

2.2. Data extraction

Included papers were imported into EndNote X9 for further analysis. The major details of measures extracted included task name, the cited source of the task (task reference), stimulus type, response type, as well as scoring method and number of raters for measures using an open-ended response format. When more than one published task was merged and scored together as one larger task without distinguishing the individual components, the combined task was considered a new task. Results of the measures were also extracted, including the maximum score possible, observed range of scores, mean score, and standard deviation of scores. Psychometric properties extracted included reliability indices and any evidence of validity, and were limited to the original psychometric properties calculated from the data collected for each study. Any modifications to the measures specified by authors were also recorded. The extraction of results and psychometric properties was limited to the subset of neurotypical adult participants.

2.3. Coding

Information about each measure is summarized in Table 1. We scored the following attributes if the criteria were met in the target paper, or if they were met in the original paper from which the ToM task was derived. The task name of each measure was unified after checking the test procedures and task references of each record. Stimulus type included stories, videos (i.e., featuring real people), photos (i.e., featuring at least a part of the faces of real people with or without context), single cartoons (i.e., single cartoon presented to prompt interpretation by participants), cartoon sequencing, animations, text in sentences, and others (e.g., interactive games). Response type included forced-choice, open-ended, sequencing, and others (e.g., pointing along a continuum). Scoring method of open-ended measures included binary scale, k-point scale (k varies from three to seven), count or proportion of certain types of response, or was not specified. The original aim of the measure first separated measures that involved testing neurotypical adults in the source paper from those that did not. For those that did, the aims were categorized into five types: population comparison (i.e., between neurotypical adults and other age groups of clinical groups), individual differences, neural underpinnings including lesion studies, experimental condition comparison, and others (e.g., norm setting). It was possible to have multiple codes for stimulus type, response type, scoring method, and original aim of measure, as the same measure may have been adapted in different ways in different studies. The tasks were coded as aiming to measure individual differences if this was explicitly stated, or if the source paper examined correlations between the task score and other behavioural or demographic variables.

Correlates were categorised into eight major types, and four subtypes: (1) traits, specifically clinical traits (e.g., autistic quotient, psychosis proneness), social traits (e.g., empathic quotient, empathic concern), personality traits (e.g., Big Five), and other traits (e.g., gender identity scale ratings); (2) social cognition measures (e.g., social intelligence, emotion recognition); (3) cognitive abilities (e.g., general intelligence, executive functions); (4) social functioning (e.g. social appropriateness, negotiation ability); (5) social outcomes (e.g. interpersonal relationship quality, intimate network size); (6) demographics; (7) miscellaneous (e.g. fatigue, fiction exposure); and (8) other ToM measures.

The mean percent of maximum possible (POMP) score for each measure was calculated by taking the average of the mean scores in all the studies adopting the measure. In cases where it was impossible to calculate the POMP score (i.e., the mean score was presented as a raw score without reporting the maximum score possible), the entry was omitted as different studies could have adopted different scoring methods and have different maximum scores possible even when using the same measure. Where number of errors were reported and the total number of trials were reported, the mean score of the study was calculated by reversing the average proportion of error to proportion of correct responses. However, we did not calculate the POMP scores for subscales of different types of errors (e.g., undermentalising and overmentalising errors in MASC), as they reflected the type of error committed by participants rather than participants' performance.

To provide an accessible summary, reliability and validity information was coded with a three-colour system, as presented in Fig. 2 (reliability) and Fig. 3 (validity). Green is the most satisfactory, followed by yellow, and red indicates caution. The information was coded on a study level, as shown in Fig. 2 and Fig. 3, and explained below. Table 2 and Table 3 show the number of studies in which the reliability or validity of each measure was coded green, yellow, and red. The full set of extracted data are available from the link at the end of this section.

For reliability, internal consistency of a measure was coded green if the Cronbach's alpha, Guttman's lambda, or omega reported in a study was .7 or above (Cortina, 1993) or intra-class correlation (ICC) was .75 or above (Fleiss, 1986); it was coded yellow if alpha/lambda/omega indices were between .6 and .7, ICC was between .5 and .75, or split-half reliability was between .5 and .75. If different indices in the same study conflicted in colour coding, the coding was decided upon the value of the alpha/lambda/omega index. Test-retest reliability was coded green if the correlation coefficient between two time points administering the same test within eight weeks was .70 or above or intra-class correlation (ICC) was .75 or above, yellow if the correlation was between .4 and .70 (.75 for ICC), and red if the correlation was below .4 (Cicchetti, 1994; Fleiss, 1986). Inter-rater reliability was coded green if the Cohen's Kappa or intra-class correlation was .75 or above (Mordal et al., 2010); average indices between .4 and .75 were coded yellow and those below .4 were coded red. An observed factor structure being consistent with the one hypothesised was taken as evidence supporting the factor

Table 1
List of measures identified (in descending order of occurrences in studies). The "top eight" measures discussed in most detail in the text are shaded.

Measure name	Task reference	No. of studies	Original aim	(Range of) mean age	Stimulus type	Response type	Item scoring method (* refers to scoring method used in the original reference; # refers to total score)	Scoring attribute
Reading the Mind in the Eyes Test	Baron-Cohen et al., 2001	149	Population comparison (clinical); individual differences	Range: 18.1-59.2 Mean: 29.0	Photos (eyes)	Forced-choice (3/4 options)	Binary scale	Correctness
Strange Stories Task	Happé, 1994	33	Population comparison (clinical)	Range: 18.6-47.7 Mean: 28.6	Stories	Open-ended	Binary scale/3-point scale	Correctness
Faux pas recognition test	Baron-Cohen et al., 1999	28	* Population comparison (clinical); task comparison (designed for children)	Range: 18.6-59.2 Mean: 32.8	Stories	Open-ended	Binary scale/3-point scale	Correctness
Hinting task	Corcoran et al., 1995	25	Population comparison (clinical)	Range: 20.1-51.7 Mean: 31.6	Stories/ Videos	Open-ended	Binary scale/3-point scale/4-point scale	Correctness
ToM Picture Stories task	Brüne, 2003	12	Population comparison (clinical)	Range: 20.5-46.3 Mean: 34.0	Cartoons (sequence)	Forced-choice (3 options)/Sequencing & Open-ended	7-point scale (sequencing); n/a (sequencing time); #23 max (open-ended questionnaire total score)	Correctness; n/a; correctness
Imposing memory test	Kinderman et al., 1998	11	Population comparison (group split by other variables)	Range: 20.3-53.0 Mean: 28.8	Stories/ Videos	Forced-choice (binary)	Binary scale	Correctness
MASC	Dziobek et al., 2006	11	Population comparison (clinical); individual differences	Range: 19.9-47.0 Mean: 28.6	Videos	Forced-choice (4 options)	Binary scale	Correctness/ (propensity if taking into consideration the type of error committed)
Animations task	Abell et al., 2000	10	Population comparison (clinical)	Range: 19.3-32.3 Mean: 24.9	Animations	Forced-choice (4 options)/Open-ended	Binary scale/3-point scale/*6-point scale (intentionality subscale)	Correctness/propensity
False belief task (1st-order + 2nd-order)	Perner & Wimmer, 1985	8	*Developmental differences (designed for children)	Range: 21.9-35.5 Mean: 27.1	Cartoons (sequence)/Stories	Forced-choice (binary)/Open-ended	Binary scale	Correctness
TASIT	McDonald et al., 2003	8	Population comparison (clinical)	Range: 19.7-40.7 Mean: 29.6	Videos	Forced-choice (binary/3 options)	Binary scale	Correctness
Yoni task	Shamay-Tsoory & Aharon-Peretz, 2007	6	Population comparison (clinical)	Range: 19.8-25.9 Mean: 22.7	Illustrated items	Forced-choice (4 options)	Binary scale	Correctness
Short Story Task	Dodell-Feder et al., 2013	5	Individual differences	Range: 19.4-27.8 Mean: 23.6	Stories	Open-ended	Binary scale (spontaneous subscale); 3-point scale (explicit mental subscale)	Correctness/propensity
Director task	Keysar et al., 2000	4	* Experimental condition comparison (age not mentioned)	Range: 19.1-23.0 Mean: 21.3	Interactive game	Action	Binary scale (error measure); n/a (RT measure)	Correctness; n/a

(*continued on next page*)

Table 1 (*continued*)

Picture sequencing task	Langdon et al., 1997	4	Population comparison (clinical)	Range: 32-47.7 Mean: 40.15	Cartoons (sequence)	Sequencing & Open-ended	5-point scale/3-point scale/not specified (sequencing); proportion of mental state terms in open-ended responses	Correctness; propensity
Reading the mind in the voice task	Golan et al., 2007	4	Population comparison (clinical); individual differences	Range: 19.3-35.6 Mean: 24.5	Audios	Forced-choice (4 options)	Binary scale	Correctness
Visual perspective taking task	Samson et al., 2010	4	Experimental condition comparison	Range: 21.7-40.9 Mean: 31.2	Pictorial probes	Forced-choice (binary)	Mean response time divided by proportion correct	Correctness
Comic strip task	Sarfati et al., 1997	3	Population comparison (clinical)	Range: 19.0-38.0 Mean: 27.4	Cartoons (sequence)	Forced-choice (3 options)	Binary scale	Correctness
Edinburgh Social Cognition Test (ESCoT)	Baksh et al., 2018	3	Individual differences	Range of means: 22.5-38.4 Mean of means: 32.8	Animations	Open-ended	4-point scale	Correctness
EmpaToM	Kanske et al., 2015	3	Neural underpinnings; individual differences	Range: 28.7-40.9 Mean: 36.8	Videos	Forced-choice (3 options)	Binary scale (score measure); n/a (RT measure)	Correctness
Moral judgment task	Young et al., 2007	3	Neural underpinnings	Range of means: 34.4-56.6 Mean of means: 41.7	Stories	Ratings	Rating differences between ToM and baseline conditions	Rating differences
Reading the mind in films task	Golan, Baron-Cohen, & Hill, et al., 2006	3	Population comparison (clinical); individual differences	Range: 35.6-38.4 Mean: 37.2	Videos	Forced-choice (4 options)	Binary scale	Correctness
Theory of mind stories task	Frith & Corcoran, 1996	3	Population comparison (clinical)	Range: 39-40.9 Mean: 39.6	Stories (with cartoons)	Open-ended	Binary scale	Correctness
Visual jokes test	Corcoran et al., 1997	3	Population comparison (clinical)	Range of means: 20.3-37.8 Mean of means: 27.0	Cartoons (single)	Open-ended	4-point scale/Binary scale	Correctness
Adult Theory of Mind test (A-ToM)	Brewer et al., 2017	2	Population comparison (clinical)	Range: 22.4-26.1 Mean: 24.3	Videos	Forced-choice (binary) & Open-ended	3-point scale/Binary scale;not applicable for RT	Correctness; RT
Attribution of intention task	Brunet, Sarfati, Hardy-Baylé & Decety, 2000	2	Neural underpinnings	Range: 30.9-47.7 Mean: 39.3	Cartoons (sequence)	Forced-choice (3 options)	Binary scale	Correctness
Cambridge mindreading face battery	Golan, Baron-Cohen & Hill, 2006	2	Population comparison (clinical)	Range: 22.2-22.5 Mean: 22.3	Videos	Forced-choice (4 options)	Binary scale	Correctness
Combined stories task	Achim et al., 2012	2	Population comparison (clinical)	Range: 24.2-25.2 Mean: 24.7	Stories	Open-ended	Binary scale/3-point scale	Correctness
False belief task (1st-order)	Wimmer & Perner, 1983	2	*Developmental differences (designed for children)	Range: 20.4-40.2 Mean: 30.3	Animations/ Cartoons (sequence)/Stories	Forced-choice (3 options)/Open-ended	Binary scale	Correctness

(*continued on next page*)

Table 1 (*continued*)

Mind Reading in Films task	Tahazadeh et al., 2020	2	Population comparison (clinical); individual differences	Range: 21.6-23.6 Mean: 22.6	Videos	Forced-choice (4 options)	Binary scale	Correctness
Modified Picture Stories-Theory of Mind Questionnaire (MPS-TOMQ)	Calso et al., 2019	2	Population comparison (age); individual differences	Range: 25.4-25.6 Mean: 25.5	Cartoons (sequence)	Sequencing & Open-ended	7-point scale (sequencing); n/a (sequencing time); not specified (TOMQ)	Correctness; n/a; not specified
Second-order false-belief task	Pickup & Frith, 2001	2	Population comparison (clinical)	Range: 32.7-33.5 Mean: 33.1	Playmobil figures/ Stories	Open-ended	3-point scale/4-point scale	Correctness
Situational test of emotion understanding	MacCann & Roberts, 2008	2	Individual differences	Range: 20.3-20.4 Mean: 20.4	Sentences	Forced-choice (5 options)	5-point scale	Not specified
Spontaneous ToM Protocol (STOMP)	Rice & Redcay, 2015	2	Neural underpinnings; individual differences	Mean: 20.3	Videos	Open-ended	Proportion of internal state statements	Propensity
Story comprehension test	Channon & Crawford, 2000	2	Lesion study	Range: 19.4-20.2 Mean: 19.8	Stories	Open-ended	3-point scale/binary scale*	Correctness(*); propensity*
Unexpected outcomes test	Dyck et al., 2001	2	*Developmental differences; individual differences (designed for children)	Range: 19.5-36.6 Mean: 28.1	Stories	Open-ended	3-point scale	Correctness
Virtual assessment of mentalising ability (VAMA)	Canty et al., 2017	2	Individual differences	Range: 25.9-45.6 Mean: 35.8	Interactive game	Forced-choice (4 options)	3-point scale/Binary scale	Correctness
Arena of Emotions Tasks	Rosenblau et al., 2015	1	Population comparison (clinical); individual differences	Mean: 32.4	Videos	Forced-choice (4 options)	Binary scale	Correctness
Attitudinal subset (APT) of the Aprosodia Battery	Orbelo et al., 2005	1	Population comparison (age)	Mean: 34.8	Audios	Forced-choice (binary)	Binary scale	Correctness
Belief-desires task	Apperly et al., 2011	1	Population comparison (age); experimental condition comparison	Mean: 20.3	Sentences	Forced-choice (binary)	n/a (RT measure)	n/a
Cartoon Reading the mind in the eyes task	Atherton, G. & Cross, L., 2021	1	Individual differences	Mean: 21.9	Cartoons (single)	Forced-choice (4 options)	Binary scale	Correctness
Cartoon stories ToM paradigm	Kosmidis, 2011	1	Population comparison (clinical); individual differences	Mean: 37.4	Cartoons (sequence)	Forced-choice (binary)	Binary scale	Correctness
Computerised false-belief task	Wang et al., 2021	1	Experimental condition comparison; individual differences	Mean: 19.5	Cartoons (sequence)	Forced-choice (binary)	n/a (RT measure)	n/a

(*continued on next page*)

Table 1 (*continued*)

Conflicting beliefs and emotions task	Shaw et al., 2004	1	* Lesion study (age not mentioned)	Mean: 30.6	Stories	Open-ended	Binary scale	Correctness
Conversations and Insinuations task	Ouellet et al., 2010	1	Population comparison (clinical)	Mean: 23.1	Videos	Forced-choice (4 options)	Binary scale	Correctness
Dewey Social Stories Test	Dewey, 1991	1	Population comparison (clinical)	Mean: 34.8	Stories	Forced-choice (4 options)	4-point scale	Deviation from most common response
Emotion Attribution task	Blair & Cipolotti, 2000	1	Lesion study	Mean: 40.2	Stories	Open-ended	Binary scale	Correctness
Faces test (Adoplhs et al.)	Adoplhs et al., 2002	1	*Lesion study (age not mentioned)	Mean: 36.6	Photos (face)	Forced-choice (binary)	Binary scale	Correctness
Faces test (Baron-Cohen et al.)	Baron-Cohen et al., 1997	1	Population comparison (clinical)	Mean: 20.7	Photos (face)	Forced-choice (binary)	Binary scale	Correctness
Irony perception task	Langdon et al., 2002	1	Population comparison (clinical)	Mean: 20.0	Stories	Forced-choice (binary)	Binary scale	Correctness
Joke-appreciation task	Happé et al., 1999	1	*Population comparison (clinical) (designed for the elderly)	Mean: 32.0	Cartoons (single)	Open-ended	4-point scale	Correctness
Judgement of preference	Girardi, MacPherson, & Abraham, 2011	1	Population comparison (clinical); experimental condition comparison	Mean: 38.4	Illustrated items	Forced-choice (4 options)	Binary scale	Correctness
Multifaceted Empathy Test	Dziobek et al., 2007	1	Population comparison (clinical)	Mean not reported	Photos (real person in context)	Forced-choice (4 options)	Binary scale	Correctness
Nonverbal cartoon task	Gallagher et al., 2000	1	Neural underpinnings	Mean: 42.0	Cartoons (single)	Open-ended	Binary scale	Correctness
Novel wisdom/ToM task	Rakoczy, H. et al., 2018	1	Population comparison (age); individual differences	Mean: 24.3	Stories	Open-ended	3-point scale	Correctness
Perspective Taking Task	Gallant, C., & Good, D., 2020	1	Population comparison (group split by other variables); individual differences	Mean: 19.8	Stories	Ratings	Average ratings for correct responses	Ratings
Pragmatic language comprehension task	Koster-Hale, Dodell-Feder, Saze, unpublished	1	n/a	Mean: 20.3	Sentences	Forced-choice (binary)	Binary scale	Not specified
Rutherford stories task	Rutherford, 2004	1	Experimental condition comparison	Mean: 24.7	Stories	Forced-choice (binary)	Binary scale	Correctness

(*continued on next page*)

Table 1 (*continued*)

Sandbox task	Sommerville et al., 2010	1	Population comparison (age); experimental condition comparison; individual differences	Mean: 37.7	Stories	Pointing to a location within a continuous space	Distance away from first location to second location	Distance away
Self-referential mentalizing interview	Ballespi, S. et al., 2019	1	Individual differences	Mean: 21.1	Interview questions	Ratings	n/a	n/a
Social Attribution Task-Multiple Choice	Klin, 2000	1	Population comparison (clinical)	Mean: 32.0	Animations	Forced-choice (4 Options)/*Open-ended (original measure)	Binary scale/*7-point scale/*Proportion of using mental state terms	Correctness/propensity
Social Cognition Screen Questionnaire (ToM subscale)	Roberts et al., 2011	1	* Individual differences (designed for clinical patients)	Mean: 37.8	Stories	Forced-choice (binary)	Binary scale	Correctness
Social stories questionnaire	Lawson, Baron-Cohen & Wheetwright, 2004	1	Population comparison (clinical)	Mean: 20.1	Stories	Forced-choice (binary)	Binary scale	Correctness
Story-Based Empathy Task	Dodich, A. et al., 2015	1	Norm setting	Mean: 49.6	Cartoons (sequence)	Forced-choice (3 options)	Binary scale (accuracy); 5-point scale (equivalent score)	Correctness; deviance from median
Strange stories film task	Murray et al., 2017	1	Population comparison (clinical); individual differences	Mean: 32.5	Videos	Open-ended	3-point scale	Correctness
Strange stories task + ToM Stories task	*Licata, M. et al., 2016 (the study that used this combined measure)	1	* n/a (refer to the two separate measures)	Mean: 38.0	Stories	Open-ended	4-point scale *(0/0.5/1/2)	Correctness
The cartoon vignette	Sebastian et al., 2012	1	Neural underpinnings	Mean: 21.3	Cartoons (sequence)	Forced-choice (binary)	Binary scale	Correctness
The situational test of emotion management	MacCann & Roberts, 2008	1	Individual differences	Mean: 20.4	Hypothetical scenarios	Forced-choice (4 options)/*Ratings (original article)	Binary scale/Weighted score (forced-choice); *distance from expert ratings (ratings)	Correctness(*); distance from expert rating*
Theory of Mind Assessment Scale (Th.o.m.a.s.)	Bosco et al., 2009	1	Population comparison (clinical)	Mean: 40.7	Interview questions	Open-ended	5-point scale	Coherence, clearness and abundance of contextualised examples
Theory of mind in dialogue	Dwyer et al., 2020	1	Population comparison (clinical)	Mean: 40.9	Interview questions	Open-ended	Number of references to own and others' beliefs	Propensity
ToM stories task	German & Hehman, 2006	1	Population comparison (age); individual differences	Mean: 38.8	Stories	Forced-choice (binary)	Binary scale	Correctness

(*continued on next page*)

Table 1 (*continued*)

ToM task (false belief + faux pas)	Henry et al., 2011	1	Population comparison (clinical)	Mean: 43.7	Stories	Open-ended	Binary (9 max) (FB1 total score); 3-point scale (FB2); 3-point scale (faux pas)	Correctness
ToM videos task (belief reasoning task)	Apperly et al., 2004	1	Lesion study	Mean: 38.8	Videos	Forced-choice (binary)	Binary scale	Correctness
ToM videos test	Sullivan & Ruffman, 2004	1	Population comparison (age); individual differences	Mean: 36.1	Videos	Forced-choice (binary)	Binary scale	Correctness
ToM-HCAT	Aykan, S. & Nalcaci, E., 2018	1	Individual differences	Mean: 21.3	Cartoons (single)	Forced-choice (4 options)	Binary scale	Correctness
Verbal stories ToM paradigm	Kosmidis, 2011	1	Population comparison (clinical); individual differences	Mean: 37.4	Stories	Open-ended	3-point scale (hinting task stories); not specified (FB1, FB2, 1st order deception, 2nd order deception)	Correctness

RT refers to response time. (Achim et al., 2012; Adolphs et al., 2002; Apperly et al., 2004; Atherton and Cross, 2022; Aykan and Nalçacı, 2018; Baksh et al., 2018; Ballespí et al., 2019; Baron-Cohen et al., 1997; Blair, 2000; Bosco et al., 2009; Brewer et al., 2017; Brunet et al., 2000; Calso et al., 2019; Channon and Crawford, 2000; Corcoran et al., 1997; Dewey, 1991; Dodell-Feder et al., 2013; Dodich et al., 2015; Dwyer et al., 2020; Dyck et al., 2001; Dziobek et al., 2008; Frith and Corcoran, 1996; Gallagher et al., 2000; Gallant and Good, 2020; Gilpin, 1993; Girardi et al., 2011; Golan et al., 2006; Golan et al., 2006; Golan et al., 2007; Happé et al., 1999; Henry et al., 2011; Kanske et al., 2015; Klin, 2000; Kosmidis et al., 2011; Koster-Hale et al., 2012; Langdon et al., 2002; Langdon et al., 1997; Lawson et al., 2004; Licata et al., 2016; MacCann and Roberts, 2008; McDonald et al., 2003; Murray et al., 2017; Orbelo et al., 2005; Ouellet et al., 2010; Perner and Wimmer, 1985; Pickup and Frith, 2001; Rakoczy et al., 2018; Rice and Redcay, 2015; Roberts et al., 2011; Rosenblau et al., 2015; Rutherford, 2004; Samson et al., 2010; Sarfati et al., 1997; Sebastian et al., 2012; Shah et al., 2017; Shamay-Tsoory et al., 2007; Shaw et al., 2004; Sommerville et al., 2013; Sullivan and Ruffman, 2004; Tahazadeh et al., 2020; Wang et al., 2021; Wang et al., 2016; Young et al., 2007).

structure of the measure. Most of the time, the measures proposed to capture a unitary ToM component, and the factor structure was supported if the results showed a good fit to a one-factor model. In other measures that included a control scale or proposed several subscales, a good fit to a two-factor model that distinguished the ToM subscale and the control subscale, or the proposed subscales, were treated as evidence for the proposed factor structures.

Validity was colour-coded based on whether the studies reported evidence for or against different kinds of validity. Green was coded when there was only supporting evidence within a single study; yellow referred to mixed evidence within a single study (i.e. having both evidence that supports and opposes validity in the same study, such as reporting one correlation larger than the effect size threshold we will later specify, and another correlation smaller than the threshold), and red was coded when there was only evidence against validity in the specific way, within a single study. We coded for four types of validity evidence, conceptually similar to convergent validity, criterion-related validity, known-group validity and discriminant validity.

We coded for "broad" convergent validity and "narrow" convergent validity. Reports of performance on the measure correlating with other social cognition or social ability measures, not limited to ToM, were taken as evidence of broad convergent validity. Positive evidence was characterised by a Pearson's or Spearman's correlation coefficient of .19 (taking the absolute value) or higher, which is the median effect size in individual differences studies (Gignac, Szodorai, 2016). By adopting this criterion, which is less stringent than Cohen's convention of .30 for a medium effect size (Cohen, 1992), we expect to err on the side of an optimistic picture of convergent validity displayed by the identified tasks. Correlations of task performance and general social abilities or relevant clinical traits, specifically autistic quotient (AQ) or alexithymia trait scores, were also included as evidence regarding broad convergent validity, for the questionnaires include components that tapped on social cognitive abilities. The same .19 threshold explained above was applied in such cases. In most cases evidence in favour of convergent validity came from positive correlations, but it was also possible for negative correlations to provide positive evidence (e.g., when one of the correlated measures examined response time (RT), or when participants' ToM performance was correlated with clinical traits associated with social difficulties). For narrow convergent validity, we investigated interrelations among the ToM tasks identified in this review for relevant evidence. Two tasks were taken as correlated in a study if there was at least one correlation that exceeded the .19 threshold between any subscales of the two tasks. Any lower correlations reported in studies were considered evidence against interrelation between two tasks.

Criterion-related validity was supported by evidence suggesting a correlation between performance on the measure and social functioning or social outcomes (e.g., interpersonal relationship quality, community functioning, social functioning scale performance). Known-group validity was supported by reports of differences in performance on the measure between the neurotypical adult control group and clinical groups showing social deficits, specifically autism spectrum disorder (ASD) and schizophrenia, or between participants grouped by high versus low autistic or schizophrenic traits, or either children or older adults. Discriminant validity was supported by results showing that (1) the measure contributed to unique variance in criterion variables including social functioning and social outcomes after controlling for at least one of three confounds: verbal ability, general intelligence, executive functions; (2) only the subscale(s) relevant to ToM but not the control subscale(s) correlated with the criterion variables; (3) known-group differences in task performance remained significant after controlling for at least one of the three confound variables; (4) known-group differences in the ToM-relevant and control subscales were dissociated; or (5) known-group differences in ToM-relevant subscale(s) remained significant after controlling for the scores on control subscale(s).

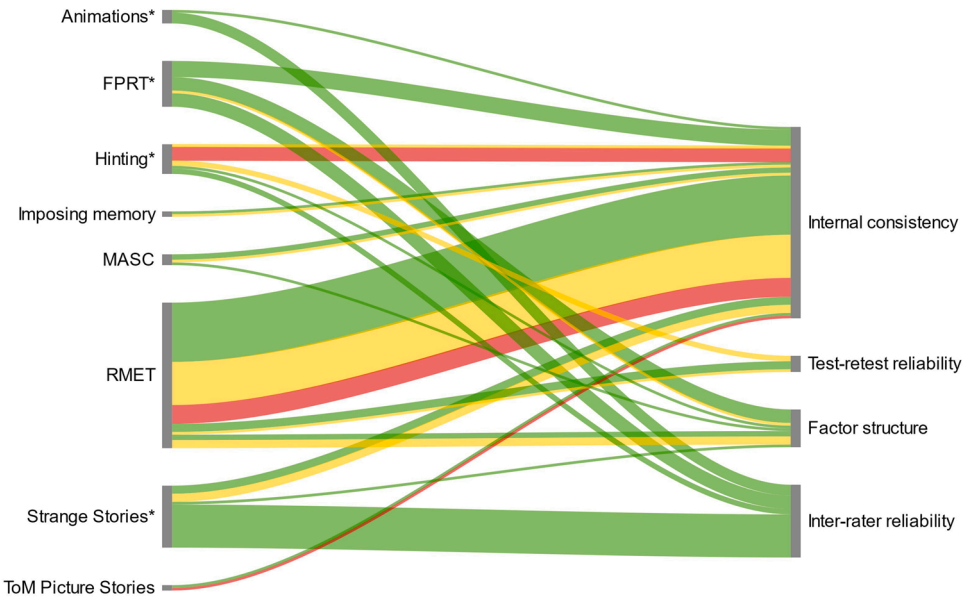

Fig. 2. Available evidence regarding reliability of the top 8 measures. The diagram depicts the availability of evidence for or against reliability of the top eight popular measures, including Animations Task (Animations*), Faux Pas Recognition Task (FPRT*), Hinting Task (Hinting*), Imposing Memory Test (Imposing Memory), Movie for the Assessment of Social Cognition Task (MASC), Reading the Mind in the Eyes Test (RMET), Strange Stories Task (Strange Stories*), and ToM Picture Stories Task (ToM Picture Stories), in alphabetical order. The tasks that were presented in an open-ended response format in at least one study were indicated with "* ". The colour coding follows the same principle as for Table 2, with green indicating the most satisfactory evidence according to standard criteria, yellow intermediate, and red the least satisfactory. Curve width is weighted by number of studies showing relevant evidence for or against reliability. Curves extended from the same measure should have equal width if the same number of studies indicate evidence for or against the specific type regarding reliability of the same measure.

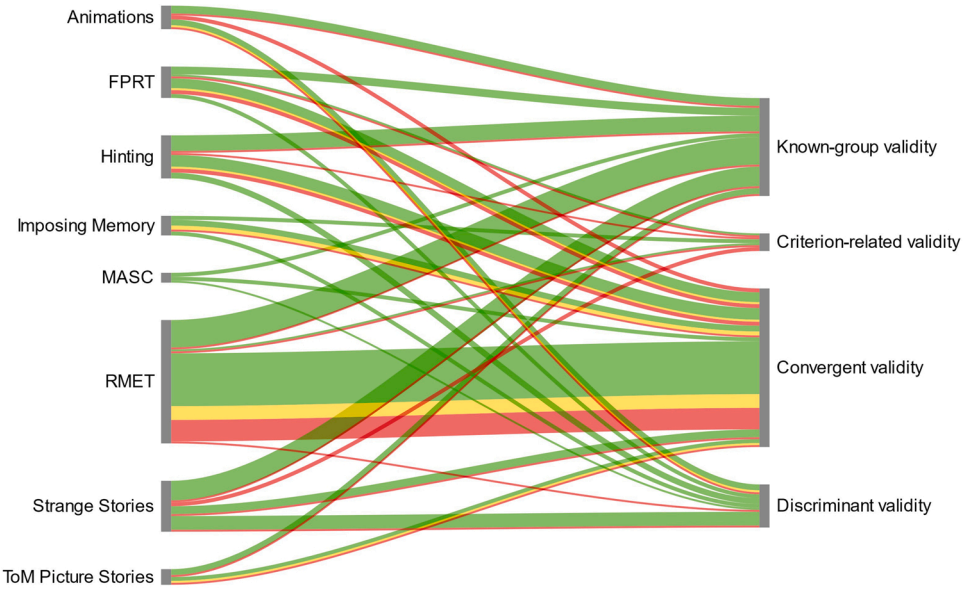

Fig. 3. Available evidence regarding validity of the top 8 measures. The diagram depicts the availability of evidence for or against validity (beyond face validity) of the top eight popular measures. The colour coding follows the same principle as for Table 3. Curve width is weighted by number of studies showing relevant evidence for or against validity. Curves extended from the same measure should be equal in width if the same number of studies indicate evidence for or against the specific type regarding validity of the same measure. Convergent validity in this diagram refers to broad convergent validity.

The full set of extracted data and the spreadsheets for coding the data are publicly available on OSF (https://osf.io/23ynq/?view_only=7f34abba115b40da99c14b1e08d97f67).

2.4. Sample characteristics

Approximately 47640 neurotypical participants aged between 18 and 65 were included in the 273 studies. The smallest study had 10 participants and the largest study included 2242 participants. The average sample size was 173 (around 62% female with all samples aggregated, excluding studies that did not report gender).

Twenty studies did not report the mean age of participants. The mean age of participants in the remaining 253 studies varied from 18.12 years to 59.27 years, and the average of mean age reported in studies was 30.04 years.

Table 2
Reliability evidence of the top 8 measures in alphabetical order (number of studies providing positive/mixed/negative evidence).

Measure name	Number of studies	Internal consistency	Test-retest reliability	Factor structure	Interrater reliability
Animations Task[a]	10	1/0/0			4/0/0
FPRT[a]	28	6/0/0		5/1/0	5/0/0
Hinting Task[a]	25	0/1/5	0/2/0	1/0/0	2/0/0
Imposing memory Test	11	1/1/0			
MASC	11	2/1/0		1/0/0	
RMET	149	22/16/7	3/1/0	2/3/0	1/1/0[b]
Strange Stories Task[a]	33	3/3/0		1/0/0	16/0/0
ToM Picture Stories Task	12	1/0/1			

[a] Tested in open-ended format in at least one study.
[b] Not tested in open-ended format but had interrater reliability reported (thus not included in the main analysis or Fig. 2).

Table 3
Validity evidence of the top 8 measures in alphabetical order (number of studies providing positive/mixed/negative evidence).

Measure name	Number of studies	Known-group validity	Criterion-related validity	(Broad) Convergent validity	Discriminant validity
Animations Task	10	4/0/1		0/0/2	3/1/1
FPRT	28	4/0/0	1/0/1	5/1/2 (2)	2/0/0
Hinting task	25	8/0/1	0/0/1	6/1/2	3/0/0
Imposing memory test	11		2/0/0	3/2/1	2/0/0
MASC	11	2/0/0		2/0/0	1/0/0
RMET	149	14/0/1	1/0/1	27/7/11 (1)	0/0/1
Strange Stories Task	33	10/0/1	0/0/2	4/0/1(1)	7/0/1
ToM Picture Stories Task	12	3/0/1		2/1/1	

The number of studies that report a relevant significance test without specifying the effect size is marked in parentheses, if applicable.

3. Results

We begin by describing the key features of the stimuli and measurement formats of the tasks identified. Next, we evaluate the psychometric properties of the tasks, with particular focus on the eight tasks for which we have the most data to inform evaluation. We also evaluate the interrelations among the measures identified.

3.1. Description of identified measures (Table 1)

We identified 75 measures that have been adopted to assess individual differences in ToM in neurotypical adults, including one unpublished measure with no further information, listed in Table 1. The mean age of participants is also summarised in Table 1. Forty-three (57%) measures were designed for detecting differences between groups in adults (e.g., adults with a known diagnosis vs. those without a diagnosis) rather than individual differences; 26 (35%) were designed to detect individual differences in adults. The mean age of participants ranged from 19.50 to 49.60 years, with an average of 30.14 years.

3.1.1. Forms of stimuli

Out of the 75 identified measures, many of the measures involved narratives or stories (52; 69%) presented as text or speech (27; 36%), videos (15; 20%), cartoon sequences (11; 15%), or animations (4; 5%). Two animation tasks featured geometric shapes rather than human agents. The forms of stimuli adopted by the remaining measures are listed in Table 1. The types of stimuli presented in five (7%) tasks were inconsistent across studies (e.g., for the Hinting task some studies presented narratives while some presented videos). How the participants were required to respond to the stimuli, and how their responses were measured, are discussed next.

3.1.2. Form of measurement

There was considerable variety in measurement methods, not only between tasks, but also when the same notional task was used in different studies. This limits the confidence with which conclusions about reliability and validity from a study using one task variant can be expected to generalise to studies using another task variant.

Response format. Most of the measures involved forced-choice responses and/or open-ended questions. Among the 75 measures, 45 (60%) involved a forced-choice between two and five alternatives, 31 (41%) required open-ended verbal responses, four (5%) involved subjective ratings (e.g., rating the likelihood of possible explanations to an agent's behaviour, or the likelihood of an agent having different emotional responses in a described social scenario), three (4%) involved picture sequencing, one (1%) involved pointing to a location within a continuous space, and one (1%) required moving a designated object as directed. Four measures (5%) involved at least two components (e.g., including both sequencing and open-ended questions). The response formats were inconsistent across studies for seven measures (9%), and one additional measure (1%) had a different number of forced-choice options in different studies.

Scoring method. As forced-choice and open-ended responses were the two most popular response formats, this subsection describes how the items were scored across different studies using the same measure. The analysis revealed considerable diversity between different methods, and between different studies using the same method.

Forced-choice measures. Dichotomous scoring that differentiated correct from incorrect answers for items was used in 39 (87%) of the forced-choice measures, while eight measures (18%) involved scoring on a k-point scale (k varies from three to seven) that rated participants' item responses according to the extent they matched with developed scoring schemes. One measure (2%) weighted scores by expert ratings of an agent's possible mental states that can arise from a described social scenario, which was collected a priori. Among the 45 measures that involved a forced-choice response format in at least one study, four (9%) have been scored using more than one of the above methods across studies.

Open-ended measures. For the 31 measures that were used with an open-ended response format in at least one study, twenty (63%) measures scored open-ended items on a k-point scale (k varies from three to seven), according to how much the participant's response matched a developed coding scheme. Fourteen measures (45%) adopted dichotomous scoring (correct or incorrect). Four measures (13%) scored participants' performance by counting or calculating the proportion of mental state references in their responses. Scoring procedures for three measures (10%) using open-ended items were not reported. Six measures (19%) were scored on more than one dimension, and 10 (32%) were scored using inconsistent methods in different studies.

Most open-ended measures were scored either according to correctness of responses, or/and evidence of a propensity to mentalise. Within the 25 (81%) open-ended measures that scored responses based on correctness, 18 (72%) scored responses on a non-binary scale and thus allowed for partial scoring. One or more of the following criteria were used to judge the score to be awarded: order of inference, extent of explicit mental state description, contextual relevance, the number of

times the experimenter gave a prompt, and explanatory power.

Seven (23%) open-ended measures captured participants' propensity to mentalise on a binary scale indicating whether the response involved mental state attribution (2 measures; 29%), a 3-, 5-, 6-, or 7-point scale reflecting the degree of deliberateness of mental state attribution (2 measures; 29%), or the occurrence of mental state references in the participants' responses in terms of count or proportion (4 measures; 57%).

Within the two measures (6%) that did not score responses on correctness or propensity, one measure scored responses on their coherence, clearness and abundance of contextualised examples; one measure did not specify the scoring criteria.

3.2. Ceiling effects and psychometric properties of measures

We first summarise the overall availability of relevant evidence from all 75 measures (see OSF for full data). Many tasks have only been used in a small number of studies, and many studies did not include evidence relevant to ceiling effects or psychometric properties. We therefore proceed to a more detailed evaluation on the eight tasks that have been used to study individual differences in neurotypical adults in 10 studies or more. As will become clear, even for these measures there is only limited evidence about reliability and validity, and we judged it even less likely that it would be possible to draw conclusions on the psychometric properties of measures where even less information was available.

3.2.1. Sensitivity to individual differences in performance

Where relevant data were available there was considerable evidence of ceiling effects. We report mean Percentage of Maximum Possible (POMP) scores and POMP score ranges to identify ceiling effects in Table 4. Table 4 shows the mean POMP scores and range of POMP scores for all measures. A task is sensitive to individual differences in a population within a particular age range when the POMP score is within the range of 20–80% (e.g., Petersen et al., 2016). We used 85% as the cut-off for indicating a ceiling effect to allow for more leniency. Measures that show a ceiling effect for at least one of the subscales are highlighted in red, including 29 measures (49% of measures that have available POMP score information) based on mean POMP score, and 13 measures (50% of measures that have available information on POMP score range) based on POMP score range. Nine measures (12%) did not have information about their mean POMP scores available because mean scores or maximum possible scores were not reported, and POMP scores were not applicable for seven measures (9%) due to their response formats (e.g., measures involving only reaction time, measures that calculated scores by taking the differences between ratings, measures that counted the number of mental state utterances). Range of POMP scores were not available for 51 (68%) measures, mostly because the measures were only used in one study.

3.2.2. Summary of reliability and validity reports

Among all 75 measures, 30 (40%) did not have information about reliability and 20 (27%) did not have information about validity (beyond face validity). Evidence of internal consistency was available from at least one study for 34 measures (45%). Evidence regarding factor structure was available for 16 (21%) measures. Only 6 (8%) measures had evidence for test-retest reliability. Evidence of inter-rater reliability was available for 16 out of 31 (52%) measures that were conducted in open-ended format in at least one study. Evidence regarding broad convergent validity was reported at least once for 49 (65%) measures, while there was evidence of known-group validity for 29 (39%) measures. Additionally, evidence of discriminant validity was available for 17 (23%) measures, and evidence regarding criterion-related validity was available for 9 (12%) measures.

Narrow convergent validity: Interrelations among measures. We examined the interrelations among ToM measures identified. Twenty-nine (39%) measures had no data bearing on their correlations with other measures. Two (4%) of 46 measures correlated with other ToM measures were not included in the analysis of this section as the correlations were not conducted specifically in the neurotypical adult group. Table 5 shows the interrelations among 44 measures (59% of 75 measures) for which there was relevant evidence, 43 of which had at least one correlation coefficient reported. When multiple correlations were conducted between different subscales or versions of the same task within the same study, we made our evaluation of positive evidence on the basis of the maximum correlation coefficient reported (taking the absolute value). This approach allowed us simplify and present the most optimistic picture of the overall correlation patterns among measures.

In total, there were 98 correlations reported, 93 (95%) of which also specified the value of the correlation coefficient. We applied a threshold of .19 for Pearson's correlation or Spearman's correlation. Out of the 93 correlations with reported coefficient values, 63 (68%) exceeded the cut-off. Among the 43 measures, 10 measures (23%) showed correlations with other measures that had an effect size smaller than the threshold.

3.2.3. The "top 8" measures

We investigated the properties of the eight tasks that were used most widely in published research. These eight measures comprised the RMET (Baron-Cohen et al., 2001; 149 studies), Strange Stories Task (Happe, 1994; 33 studies), Faux Pas Recognition Task (FPRT; Baron-Cohen et al., 1999; 28 studies), Hinting Task (Corcoran et al., 1995; 25 studies), ToM Picture Stories Task (Brüne, 2003; 12 studies), Movie for the Assessment of Social Cognition Task (MASC; Dziobek et al., 2006; 11 studies), Imposing Memory Test (Kinderman et al., 1998; 11 studies), and Animations Task (Abell et al., 2000; 10 studies). Even among these tasks, reporting of information related to reliability and validity was infrequent. The highest rate was 16 out of 33 studies employing the Strange Stories task reporting inter-rater reliability, and rates were generally much lower (Table 2). Consequently, the data available to evaluate reliability and validity is limited, and comes disproportionately from one task, the RMET. This is important to keep in mind when evaluating the summary diagrams in Figs. 2 and 3.

For the "top 8" measures, ceiling effects were shown in participants' average performance on three tasks: Strange Stories Task, FPRT, and both components of ToM Picture Stories Task as well as its total score. The minimum POMP score reported for the total score on the ToM Picture Stories Task (89.42% among six studies) also exceeded the 85% cut-off.

Table 2 and Fig. 2 list the eight measures and the availability of information on their reliability, in alphabetical order. It should be noted that information about inter-rater reliability is only available for measures that have been used with an open-ended format in at least one study, including Animations Task, FPRT, Hinting Task, and Strange Stories Task. It was noted that inter-rater reliabilities of RMET were reported in two studies in which the tasks were presented in a forced-choice format, but we do not include this information in the current summary because reports of inter-rater reliability of forced-choice measures are not informative. There was evidence regarding internal consistency for all eight measures. Table 6 shows the average Cronbach's alpha of the top eight measures, and the Hinting task is the only task that had an average Cronbach's alpha falling below .60. Five tasks had evidence for factor structure, whereas evidence regarding test-retest reliability was only available for the Hinting Task and the RMET, and this evidence was mixed.

Table 3 and Fig. 3 list the eight measures that have been adopted in 10 studies or more and the availability of information regarding their validity. All eight measures had evidence regarding broad convergent validity. Positive evidence was most frequent, but evidence was mixed for 6 of 8 tasks and only negative for one (Animations Task). We extended our analysis of narrow convergent validity to the calculation of interrelations among these eight measures by applying correction for

Table 4
POMP score of the 75 identified measures (in alphabetical order). Measures showing evidence of ceiling effects are highlighted in red. The "top eight" most frequently-used measures discussed in most detail in the text are shaded in gray.

Measure name	Stimulus type	Number of studies	Mean POMP score for neurotypical adults	POMP score range
Adult Theory of Mind test (A-ToM)	Videos	2	87.25%	n/a
Animations task	Animations	10	Appropriateness: 64.85% (7 studies) Feelings: 51.76% (2 studies) Intentionality: 66.2% (1 study)	Appropriateness: 41.13%-75.75% Feelings: 49.13%-54.38% Intentionality: n/a
Arena of Emotions Tasks	Videos	1	Indirect: 68% Direct: 67%	n/a
Attitudinal subset (APT) of the Aprosodia Battery	Audios	1	Not reported	n/a
Attribution of intention task	Cartoons (sequence)	2	84.43% (1 study)	n/a
Belief-desires task	Sentences	1	n/a	n/a
Cambridge mindreading face battery	Videos	2	75.59%	72.00%-79.18%
Cartoon Reading the mind in the eyes task	Cartoons (single)	1	67.00%	n/a
Cartoon stories ToM paradigm	Cartoons (sequence)	1	82.41%	n/a
Combined stories task	Stories	2	1st order: 93.33% (1 study) 2nd order: 83.85% (1 study)	n/a
Comic strip task	Cartoons (sequence)	3	88.80% (2 studies)	82.96%-94.64%
Computerised false-belief task	Cartoons (sequence)	1	n/a	n/a
Conflicting beliefs and emotions task	Stories	1	1st order belief: 98.00% 2nd order belief: 96.50% 1st order emotion: 89.25% 2nd order emotion: 92.50%	n/a
Conversations and Insinuations task	Videos	1	73.80%	n/a
Dewey Social Stories Test	Stories	1	92.42%	n/a
Director task	Interactive game	4	Ambiguous experimental trials: 96.80% (2 studies) Relational experimental trials: 58.00% (1 study)	Ambiguous trials: 95.00%-98.60%
Edinburgh Social Cognition Test (ESCoT)	Animations	3	Cognitive ToM: 74.18% (2 studies) Affective ToM: 88.18% (2 studies)	Cognitive ToM: 73.00%-75.37% Affective ToM: 86.93%-89.43%
Emotion Attribution task	Stories	1	90.43%	n/a
EmpaToM	Videos	3	80.48% (2 studies)	71.61%-89.35% (2 studies)
Faces test (Adoplhs et al.)	Photos (face)	1	Not reported	n/a
Faces test (Baron-Cohen et al.)	Photos (face)	1	Not reported	n/a
False belief task (1st-order + 2nd-order)	Cartoons (sequence)/Stories	8	1st + 2nd order: 91.12% (3 studies) 1st order: 90.77% (3 studies) 2nd order: 73.46% (3 studies)	1st + 2nd order: 84.89%-94.99% (3 studies) 1st order: 86.30%-95.00% (3 studies) 2nd order: 65.00%-89.57% (3 studies)
False belief task (1st-order)	Animations/Cartoons (sequence)/Stories	2	87.97%	75.93%-100%
Faux pas recognition test	Stories	28	85.90% (20 studies)	69.90%-96.00% (20 studies)
Hinting task	Stories/Videos	25	81.31% (21 studies)	62.19%-93.05% (21 studies)
Imposing memory test	Stories/Videos	11	82.44% (5 studies)	74.40%-84.13% (5 studies)
Irony perception task	Stories	1	Hit: 78.00% False alarm: 20.00% Sensitivity: 87.00%	n/a
Joke-appreciation task	Cartoons (single)	1	55.33%	n/a
Judgement of preference	Illustrated items	1	Not reported	n/a
MASC	Videos	11	Total correct: 73.57% (8 studies) Cognitive: 77.77% (2 studies) Affective: 76.45% (2 studies)	Total correct: 59.09%-78.42% (8 studies) Cognitive: 76.65%-78.89% (2 studies) Affective: 75.56%-77.33% (2 studies)
Mind Reading in Films task	Videos	2	64.89%	59.96%-69.81%

(*continued on next page*)

Table 4 (*continued*)

Modified Picture Stories-Theory of Mind Questionnaire (MPS-TOMQ)	Cartoons (sequence)	2	MPS: 85.81% (1 study) TOMQ: 55.82%	MPS: n/a TOMQ: 44.64%-67.00%
Moral judgment task	Stories	3	n/a	n/a
Multifaceted Empathy Test	Photos (real person in context)	1	Not reported	n/a
Nonverbal cartoon task	Cartoons (single)	1	97.27%	n/a
Novel wisdom/ToM task	Stories	1	90.90%	n/a
Perspective Taking Task	Stories	1	n/a	n/a
Picture sequencing task	Cartoons (sequence)	4	86.39%	82.33%-92.00% (3 studies)
Pragmatic language comprehension task	Sentences	1	Pragmatic inference accuracy: 81.90%	n/a
Reading the mind in films task	Videos	3	64.09% (1 study)	n/a
Reading the Mind in the Eyes Test	Photos (eyes)	149	Total: 72.00% (125 studies) Positive: 70.73% (7 studies) Neutral: 69.89% (7 studies) Negative: 71.36% (7 studies)	Total: 57.84%-86.12% (125 studies) Positive: 64.92%-82.00% (7 studies) Neutral: 62.50%-75.00% (7 studies) Negative: 60.00%-85.72% (7 studies)
Reading the mind in the voice task	Audios	4	71.00% (3 studies)	64.00%-78.00% (3 studies)
Rutherford stories task	Stories	1	Unweighted score: 90.00%	n/a
Sandbox task	Stories	1	n/a	n/a
Second-order false-belief task	Playmobil figures/Stories	2	57.75% (1 study)	n/a
Self-referential mentalizing interview	Interview questions	1	n/a	n/a
Short Story Task	Stories	5	Mental state reasoning: 50.17% (3 studies) Total: 63.71% (2 studies) Spontaneous mental state reasoning: 19.00% (1 study)	Mental state reasoning: 38.69%-58.06% (3 studies) Total: 59.22%-68.19% (2 studies)
Situational test of emotion understanding	Sentences	2	Not available	n/a
Social Attribution Task-Multiple Choice	Animations	1	80.95%	n/a
Social Cognition Screen Questionnaire (ToM subscale)	Stories	1	84.30%	n/a
Social stories questionnaire	Stories	1	Subtle utterances: 29.10% Blatant utterances: 57.50% Non-existence utterances: 92.15%	n/a
Spontaneous ToM Protocol (STOMP)	Videos	2	30.11%	29.11%-39.10%
Story comprehension test	Stories	2	65.50%	65.00%-66.00%
Story-Based Empathy Task	Cartoons (sequence)	1	Total: 87.39% Intention attribution: 89.33% Emotion attribution: 87.00%	n/a
Strange stories film task	Videos	1	Intention: 80.21% Mental state talk: 49.38% Interaction: 72.71%	n/a
Strange Stories Task	Stories	33	87.37% (25 studies)	55.00%-99.50% (25 studies)
Strange stories task + ToM Stories task	Stories	1	63.85%	n/a
TASIT	Videos	8	Part 2: 88.68% (4 studies) Part 3: 84.87% (7 studies)	Part 2: 84.42%-91.80% (4 studies) Part 3: 83.20%-86.70% (7 studies)
The cartoon vignette	Cartoons (sequence)	1	Affective ToM: 86.50% Cognitive ToM: 91.94%	n/a

(*continued on next page*)

Table 4 (*continued*)

The situational test of emotion management	Hypothetical scenarios	1	Not reported	n/a
Theory of Mind Assessment Scale (Th.o.m.a.s.)	Interview questions	1	First-person ToM: 95.50% Third-person allocentric ToM: 92.50% Third-person egocentric: 92.75% Second-order ToM: 91.50%	n/a
Theory of mind in dialogue	Interview questions	1	n/a	n/a
Theory of mind stories task	Stories (with cartoons)	3	Total: 90.15% (1 study)	n/a
ToM Picture Stories task	Cartoons (sequence)	12	Total: 91.63% (6 studies) Sequencing: 86.94% (5 studies) Questionnaire: 92.45% (5 studies)	Total: 89.42%-94.34% (6 studies) Sequencing: 70.00%-94.44% (5 studies) Questionnaire: 81.86%-95.83% (5 studies)
ToM stories task	Stories	1	75.29%	n/a
ToM task (false belief + faux pas)	Stories	1	n/a	n/a
ToM videos task (belief reasoning task)	Videos	1	87.39%	n/a
ToM videos test	Videos	1	88.08%	n/a
ToM-HCAT	Cartoons (single)	1	70.72%	n/a
Unexpected outcomes test	Stories	2	60.75% (1 study)	n/a
Verbal stories ToM paradigm	Stories	1	Hinting: 92.17% 1st order false belief: 97.50% 2nd order false belief: 80.00% 1st order deception: 96.00% 2nd order deception: 90.00%	n/a
Virtual assessment of mentalising ability (VAMA)	Interactive game	2	Cognitive: 66.68% (frequency); 72.65% (cumulative; 1 study) Affective: 61.50% (frequency); 69.93% (cumulative; 1 study) Total: 62.50% (frequency; 1 study)	Cognitive: 64.35%-69.00% (frequency) Affective: 60.65%-62.35% (frequency)
Visual jokes test	Cartoons (single)	3	58.00%	55.00%-66.25%
Visual perspective taking task	Pictorial probes	4	n/a	n/a
Yoni task	Illustrated items	6	Total: 92.86% (1 study) Affective: 89.62% (3 studies) Cognitive: 87.33% (3 studies)	Affective: 84.35%-92.55% Cognitive: 83.10%-90.44%

attenuation, to reduce the potential underestimation of interrelationships stemming from the measures' less-than-perfect internal consistency. This correction was possible for the top eight measures as reported values of Cronbach's alpha were available and could be averaged for each measure (see Table 6). Twenty-seven (93%) of the 29 correlations between the top eight measures had correlation coefficients reported, 18 (62%) and 21 (78%) of which exceeded the threshold of .19 before and after the correction, respectively. Table 7 lists the correlation coefficients among the top eight measures, and the number of studies that reported at least one relevant correlation that exceeded the .19 threshold, before and after correction of attenuation.

Seven out of eight tasks have some evidence regarding discriminant validity. Most of this evidence was positive, though at low frequencies. The number of studies providing evidence relevant to criterion-related validity of these measures was especially limited, with only 9 studies, and only 4 of these providing positive evidence. Notably there was no evidence regarding criterion-related validity for the Animations task, the MASC, or the ToM Picture Stories Task.

4. Discussion

The current systematic review considered measures that have been used to examine individual differences in ToM in neurotypical adults, specifically identifying the basic characteristics of the tasks, and examining ceiling effects, reliability and validity of the measures, employing a systematic strategy. We evaluated the measures with reference to established psychometric criteria, and observed that no current measure provided strong, consistent evidence of robust psychometric properties. We summarise these findings below, compare the identified measures with ToM measures for young children, make recommendations for the conduct and reporting of future research using existing measures, and identify the need to further examine psychometric properties of existing research and develop new measures that are more likely to show good psychometric properties.

4.1. Description of identified measures and standardisation of administration

Only one-third of the identified measures were specifically designed to study individual differences. Of course, tasks designed for other purposes may nonetheless succeed in measuring individual differences, but this cannot be taken for granted, and the high proportion of tasks designed for other purposes may explain evidence of poor psychometric properties. Most of the tasks employed a forced-choice response format. Open-ended responses were also common, but inter-rater reliability was not consistently reported. Moreover, while most tasks focused on scoring

the correctness of responses, a few assessed participants' propensity to make mental state attributions irrespective of correctness. This observation suggests a lack of consensus about how to operationalise individual differences in ToM. It is currently unclear whether there might truly be multiple sources of individual differences in ToM, or just incidental variation in methods.

The tasks varied in terms of stimuli and measurement formats, and tasks that were notionally the same were often implemented with different stimuli or scoring criteria between studies. While each individual study can nonetheless be evaluated on its own merits, these inconsistencies complicate the comparison of participants' performance between studies or measures. It also means that the psychometric properties of an adapted task cannot be inferred from other studies using the original version of the task (nor vice versa). For example, drawing from research on young children, research by Hughes et al. (2000) showed that the good test-retest reliability of standard false beliefs tasks was masked by the nonstandard approach of administration by Mayes et al. (1996). Similar effects are plausible in testing neurotypical adults as well.

4.2. Inspection of ceiling effects and psychometric properties

Psychometric theory provides criteria for evaluating reliability and validity, which bear on the ability of a test to measure a psychological construct (e.g., Rust, Kosinski, & Stillwell, 2021). For research on individual differences, tests must be sensitive to variation without evidence of ceiling and floor effects. A test must also show internal reliability (whereby a participant who performs well on one item tends also to perform well on other items measuring the same construct), without which it is unclear that test scores are informative about any underlying construct. It is also highly desirable that a participant who performs well on one occasion is also likely to perform well if tested later (i.e., the test shows test-retest reliability), because this indicates stability in how well the test captures the underlying construct over repeated measures. It is, of course, possible to have a highly reliable test that shows low validity because it fails to test the intended psychological construct. To evaluate validity, it is common to consider whether a test correlates with other tests of the same construct, whether it correlates with tests of other abilities, behaviours, or outcomes relevant to the construct, and whether the test is sensitive to differences between groups that differ in those abilities, behaviours or outcomes. It is also important to distinguish what a test measures from other distinct but relevant constructs. We will summarise our findings against each of these criteria.

4.2.1. Ceiling effects

Around half of the tasks showed a ceiling effect for at least one subscale (as evidenced through percentage of maximum possible scores), indicating that many tasks did not generate enough variance to study individual differences in neurotypical adults effectively. Adopting such measures can lead to erroneous conclusions that there are no individual differences in ToM in adults due to the insensitivity of the measure rather than the absence of meaningful differences in the underlying ability (e.g., Anastasi, 1948). When there is little variance within the sample, the limited spread of unique values makes it harder to detect relationships between participants' performance on the measure and other variables. While techniques for correcting range restrictions can help mitigate the underestimation of correlations with other variables, other issues, such as skewed distributions of scores, still exist, which might provide a distorted picture of the relationship between task performance and other variables of interest. Thus, ToM measures with marked ceiling effects in a target population (i.e., where the average score is $>$ 80% of maximum possible score) are unsuitable for measuring individual differences (e.g., Petersen et al., 2016).

4.2.2. Reliability and validity

Information on reliability and validity was often not reported, even among the eight ToM tasks that were adopted most frequently. Available data showed that seven out of the top eight tasks had at least acceptable internal consistency (the Hinting task was the exception). This provided support for the claim that the items in a given task reliably captured a single construct (i.e., ToM). A point to note is that good internal consistency of a task does not preclude that items vary in difficulty, or that success requires participants to adapt their reasoning to the context of individual items, as items are expected to be correlated with one another if they capture the same underlying construct. Apart from internal consistency, there was also mixed but acceptable evidence supporting inter-rater reliability and factor structure. However, very few tasks had information on test-retest reliability. If we assume that ToM is a stable trait, examining test-retest reliability is important to show that the task is tapping on the construct rather than a state that varies over time (Matheson, 2019).

As for validity, known-group validity and discriminant validity were generally satisfactory for the top eight tasks, with the exception that there was no reported evidence for known-group validity and discriminant validity for the Imposing Memory test and the ToM Picture Stories task, respectively. There was more abundant evidence regarding convergent validity for the top eight tasks, but the evidence was mixed for six tasks (except for the MASC and the Animations task). There was only evidence that support good convergent validity of the MASC, but there was no evidence for good convergent validity of the Animations task. There was especially limited information about criterion-related validity of the measures. This is a striking limitation of current literature, which means that, whether or not current tasks are measuring ToM reliably, there is little evidence (positive or negative) that they are measuring something that "matters" for social behaviour, mental health, or wellbeing.

Unsurprisingly, there was more information available regarding psychometric properties of tasks that are more frequently used. It is imperative to establish psychometric properties first, such that researchers have enough information to make informed decisions. For example, the RMET, being the most frequently used measure, had the most evidence for evaluating its psychometric properties. However, results showed that it did not exhibit the best reliability or validity. This can be because the small number of studies that adopted other tasks exaggerated the appearance of consistent evidence. Nevertheless, some tasks may demonstrate strong psychometric properties, yet lack sufficient supporting evidence due to their infrequent use. What is needed is consistent reporting of psychometric properties to generate a larger evidence base. It is suggested that researchers refer to existing guidelines on reporting psychometric properties of measures, for example, The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

Interrelations among measures. We examined the interrelations among the identified measures to investigate convergent validity. We found inconsistent evidence of intercorrelation, and some measures were not correlated with any other measures. This may reflect ceiling effects and unsatisfactory reliability of some measures, but also the possibility that ToM may be multi-dimensional rather than unidimensional. In the case of problematic ceiling effects, applying correction for attenuation to the interrelations among the top eight tasks did not change the overall picture, as only three correlations that fell below the.19 threshold before correction exceeded the threshold after correction. This observation implies that the lack of interrelations among tasks cannot be fully attributed to reliability issues. Another possible reason for the mixed interrelations is range restriction due to limited variance in task performance, as explained above in the discussion of ceiling effects, which might have masked genuine underlying associations among the tasks (Mendoza and Mumford, 1987); range restriction can also occur when some samples are highly homogenous, for instance, when assessing only university undergraduates within a single sample. This could also be a reason why we found mixed evidence

Table 5
Interrelations among identified ToM measures (in alphabetical order). Tasks that did not show any correlation with other measures with an effect size larger than the.19 threshold are highlighted in red.

Task name	Correlated task	Number of studies	Correlation index range	Number of studies reporting $r \geq .19$ (n/a)	Number of studies reporting significant correlation
Adult Theory of Mind test (A-ToM)	Animations task	1	.12-.17	0	0
	Strange Stories Task	1	.50	1	1
Animations task	Adult Theory of Mind test (A-ToM)	1	.12-.17	0	0
Arena of Emotions Tasks	RMET	1	.303-.417	1	1
Belief-desires task	Imposing memory test	1	.048	0	0
	Pragmatic language comprehension task	1	.056	0	0
	RMET	1	.115	0	0
	Spontaneous ToM Protocol (STOMP)	1	-.023	0	0
Cartoon stories ToM paradigm	Verbal stories ToM paradigm	1	.008-.529	1	1
Combined stories task	Comic strip task	1	.08	0	0
Comic strip task	Combined stories task	1	.08	0	0
Director task	Visual perspective taking task	1	-.18	0	1
Dewey Social Stories Test	Faux pas recognition test	1	-.276	1	1
	RMET	1	-.143	0	0
Edinburgh Social Cognition Test (ESCoT)	Judgement of preference	1	not reported	0 (1)	0
	Reading the mind in films task	1	.36-.42	1	1
	RMET	2	.25-.48	2	2
	Visual perspective taking task	1	-.07 - -.34	1	1
Emotion Attribution task	RMET	1	.43	1	1
	Strange Stories Task			1	1
	ToM Picture Stories task	1	.46	1	1
EmpaToM	Visual perspective taking task	1	.17	0	1
Faces test (Baron-Cohen et al.)	RMET	1	.29	1	1
	Reading the mind in the voice task	1	.22	1	1
False belief task (1st-order + 2nd-order)	RMET	1	.12	0	0
False belief task (1st-order)	RMET	1	.12	0	0

(*continued on next page*)

Table 5 (*continued*)

Faux pas recognition test	Dewey Social Stories Test	1	-.276	1	1
	RMET	5	.13-.407	4	4
	Strange Stories Task	2	.11; not reported	0 (1)	0
	ToM Picture Stories task	1	.18	0	1
	Virtual assessment of mentalising ability (VAMA)	1	.04-.45	1	1
Hinting task	Imposing memory test	2	.21	2	2
	RMET	3	.097-.28	2	2
	Second-order false-belief task	1	.201-.276	1	1
	Situational test of emotion understanding	2	.30-.33	2	2
	Social Attribution Task-Multiple Choice	1	.117	1	0
	TASIT	1	.25	1	1
	The situational test of emotion management	1	.22	1	1
	ToM Picture Stories task	1	.146	0	0
	Virtual assessment of mentalising ability (VAMA)	1	.05-.36	1	1
	Visual jokes test	1	Kendall's tau=.05 (transformed r=0.078 (Gilpin, 1993))	1	0
Imposing memory test	Belief-desires task	1	.048	1	0
	Hinting task	2	.21	2	2
	Pragmatic language comprehension task	1	-.051	1	1
	RMET	6	-.069-.42	4	4
	Situational test of emotion understanding	2	.44-.48	2	2
	Spontaneous ToM Protocol (STOMP)	2	.125-.28	1	1
	The situational test of emotion management	1	.39	1	1
Judgement of preference	Edinburgh Social Cognition Test (ESCoT)	1	not reported	0 (1)	0
	Reading the mind in films task	1	not reported	0 (1)	0
	RMET	1	not reported	0 (1)	0
MASC	RMET	1	.30	1	1
	Self-referential mentalizing interview	1	not reported; .25	1	1
Mind Reading in Films task	RMET	1	.56	1	1
Perspective Taking Task	RMET	1	-.007-.256	1	1
Picture sequencing task	Theory of mind stories task	1	.55-.63	1	1

(*continued on next page*)

Table 5 (*continued*)

Pragmatic language comprehension task	Belief-desires task	1	.056	0	0
	Imposing memory test	1	-.051	0	0
	RMET	1	.068	0	0
	Spontaneous ToM Protocol (STOMP)	1	.015	0	0
Reading the mind in films task	Edinburgh Social Cognition Test (ESCoT)	1	.36-.42	1	1
	Judgement of preference	1	not reported	0 (1)	0
	RMET	2	.38-.62	2	2
RMET	Arena of Emotions Tasks	1	.303-.417	1	1
	Belief-desires task	1	.115	0	0
	Dewey Social Stories Test	1	-.143	0	0
	Emotion Attribution task	1	.43	1	1
	Edinburgh Social Cognition Test (ESCoT)	2	.25-.48	2	2
	Faces test (Baron-Cohen et al.)	1	.29	1	1
	False belief task (1st-order + 2nd-order)	1	.12	0	0
	False belief task (1st-order)	1	.12	0	0
	Faux pas recognition test	5	.13-.407	4	4
	Hinting task	3	.097-.28	2	2
	Imposing memory test	6	-.069-.42	4	4
	Judgement of preference	1	not reported	0 (1)	0
	MASC	1	.30	1	1
	Mind Reading in Films task (Tahazadeh et al.)	1	.56	1	1
	Perspective Taking Task (scenarios from Hynes et al.)	1	-.007-.256	1	1
	Pragmatic language comprehension task	1	.068	0	0
	Reading the mind in films task	2	.38-.62	2	2
	Reading the mind in the voice task	1	.35	1	1
	Short Story Task (Dodell-Feder et al.)	4	.18-.42	3	4

(*continued on next page*)

Table 5 (*continued*)

	Situational test of emotion understanding	2	.53-.54	2	2
	Social Attribution Task-Multiple Choice	1	.331	1	1
	Spontaneous ToM Protocol (STOMP)	2	-.16 - -.115	0	0
	Strange Stories Task	4	.14-.42; not reported	2 (1)	1
	TASIT	1	.371	1	1
	The situational test of emotion management	1	.42	1	1
	ToM Picture Stories task	2	.43-.535	2	2
	Unexpected outcomes test	1	.26	1	1
	Yoni task	1	.26	1	1
Reading the mind in the voice task	Faces test (Baron-Cohen et al.)	1	.22	1	1
	RMET	1	.35	1	1
Second-order false-belief task	Hinting task	1	.201-.276	1	1
Self-referential mentalizing interview	MASC	1	not reported; .25	1	1
Short Story Task	RMET	4	.18-.42	3	4
Situational test of emotion understanding	Hinting task	2	.30-.33	2	2
	Imposing memory test	2	.44-.48	2	2
	RMET	2	.53-.54	2	2
	The situational test of emotion management	1	.62	1	1
Social Attribution Task-Multiple Choice	Hinting task	1	.117	0	0
	RMET	1	.331	1	1
Spontaneous ToM Protocol (STOMP)	Belief-desires task	1	-.023	0	0
	Imposing memory test	2	.125 - .28	1	1
	Pragmatic language comprehension task	1	.015	0	0
	RMET	2	-.16 - -.115	0	0
Strange Stories Task	Adult Theory of Mind test (A-ToM)	1	.50	1	1
	Emotion Attribution task	1	.69	1	1
	ToM Picture Stories task	1	.42	1	1
	Faux pas recognition test	2	.11; not reported	0 (1)	0
	RMET	4	.14-.42; not reported	2 (1)	1

(*continued on next page*)

Table 5 (*continued*)

TASIT	Hinting task	1	.25	1	1
	RMET	1	.371	1	1
	ToM Picture Stories task	1	.525	1	1
The situational test of emotion management	Hinting task	1	.22	1	1
	Imposing memory test	1	.39	1	1
	RMET	1	.42	1	1
	Situational test of emotion understanding	1	.62	1	1
Theory of mind stories task	Picture sequencing task	1	.55-.63	1	1
ToM Picture Stories task	Emotion Attribution task	1	.46	1	1
	Faux pas recognition test	1	.18	0	1
	Hinting task	1	.146	0	0
	RMET	2	.43-.535	2	2
	Strange Stories Task	1	.42	1	1
	TASIT	1	.525	1	1
Unexpected outcomes test	RMET	1	.26	1	1
Verbal stories ToM paradigm	Cartoon stories ToM paradigm	1	.008-.529	1	1
Virtual assessment of mentalising ability (VAMA)	Faux pas recognition test	1	.04-.45	1	1
	Hinting task	1	.05-.36	1	1
	Yoni task	1	.01-.21	1	0
Visual jokes test	Hinting task	1	Kendall's tau=.05 (transformed r=.078 (Gilpin, 1993))	0	0
Visual perspective taking task	Director task	1	-.18	0	1
	Edinburgh Social Cognition Test (ESCoT)	1	- .34 - -.07	1	1
	EmpaToM	1	.17	0	1
Yoni task	RMET	1	.26	1	1
	Virtual assessment of mentalising ability (VAMA)	1	.01-.21	1	0

for broad convergent and criterion-related validity of tasks that exhibited ceiling effects. The lack of interrelations among certain tasks might also be attributed to attenuation of correlations due to distinct task demands for different tasks. A latent variable approach is one way of addressing this problem of task impurity: if a common latent factor emerges this provides evidence that the tasks capture a common construct despite having different incidental requirements.

Moreover, the inconsistency of interrelations among tasks might reflect multidimensionality of ToM. ToM is a loosely defined construct with diverse operationalisations (Apperly, 2010; Happé et al., 2017; Schaafsma et al., 2015; Warnell and Redcay, 2019). While all tasks reviewed had face validity as ToM tasks, researchers need to look

Table 6
Average Cronbach's alpha of the top 8 measures (in alphabetical order).

Task name	Average Cronbach's alpha	Number of reports
Animations task	0.80	1
Faux pas recognition test	0.87	7
Hinting task	0.55	6
Imposing memory test	0.86	1
MASC	0.76	3
Reading the Mind in the Eyes Test	0.68	37
Strange Stories Task	0.68	5
ToM Picture Stories task	0.65	2

beyond face validity, because superficial resemblance to the construct of interest does not guarantee accurate and specific assessment. For example, despite the face validity of the RMET there is evidence that this task measures emotion perception rather than theory of mind (Oakley et al., 2016). This issue particularly warrants concern when considering that different tasks require participants to engage in different activities, including but not limited to making mental state inferences about characters from vignettes, photos and videos, interpreting non-literal speech, and recognising social transgressions. Face validity does not elucidate whether a task in fact captures a common underlying construct. While studies using latent variable analysis have identified a single underlying latent construct of ToM in early childhood, middle childhood and adolescence (e.g., Devine et al., 2023; Hughes, Devine, & Wang, 2018), similar work with adults has yet to be undertaken.

Another possible reason for inconsistent associations among tasks is that some tasks may not index ToM ability. It is difficult to establish if a task captures ToM or not when researchers have not mapped out the taxonomy of abilities that make up the construct of ToM. Some literature has suggested useful theoretical principles to distinguish whether a task captures ToM, for example, the necessity to represent mental states and distinguishing one's own mental states from that of others (Quesque and Rossetti, 2020). However, tasks that fulfil such criteria might be measuring only a specific sub-ability of self-other distinction under the general latent construct of ToM, which might include motivational as well as structural components. Therefore, it is imperative for ToM researchers to tackle theoretical issues regarding the nature of ToM in adults.

4.3. Use of measures of ToM for children

In the current review we observed that tasks designed for testing developmental differences or individual differences in young children show ceiling effects in adults. It should not be surprising that tasks designed to test basic possession of mental state concepts – such as false belief tasks – show little variation in performance among participants who are far older than the age at which children typically pass these tasks. This is supported by our findings, which suggest that these tasks should not be used to study individual differences in adults.

A substantial number of the studies reviewed here adopted tasks originally designed to be "advanced" tests of ToM in older children and adolescents. These tasks are sometimes also more naturalistic, bearing higher resemblance to reality where using ToM is more complex and dynamic, compared to laboratory tasks that only focus on specific mental state concepts. Two measures designed for older children, the FPRT and unexpected outcome test, showed different results. The FPRT exhibited a ceiling effect, while the unexpected outcome test did not, although the POMP score calculated for the latter was based on just one

Table 7
Interrelations among top 8 measures before and after correction for attenuation (in alphabetical order).

Task	Correlated task	Range of *r*	Range of corrected *r*	Number of studies with uncorrected $r \geq .19$ (n/a)	Number of studies with corrected $r \geq .19$ (n/a)	Number of reports
Faux pas recognition test	Reading the Mind in the Eyes Test	.13 – .41	.17 – .53	4	4	5
	Strange Stories Task	.11; not reported	.14; not reported	0 (1)	0 (1)	2
	ToM Picture Stories task	.18	.24	0	1	1
Hinting task	Imposing memory test	.21	.31	2	2	2
	Reading the Mind in the Eyes Test	.10 – .28	.16 – .46	2	2	3
	ToM Picture Stories task	.15	.25	0	1	1
Imposing memory test	Hinting task	.21	.31	2	2	2
	Reading the Mind in the Eyes Test	-.07 – .42	-.09 – .55	4	4	7
MASC	Reading the Mind in the Eyes Test	.30	.42	1	1	1
Reading the Mind in the Eyes Test	Faux pas recognition test	.13 – .407	.17 – .53	4	4	5
	Hinting task	.10 – .28	.16 – .46	2	2	3
	Imposing memory test	-.07 – .42	-.09 – .55	4	4	7
	MASC	.30	.42	1	1	1
	Strange Stories Task	.14 – .42; not reported	.21 – .62; not reported	2 (1)	3 (1)	4
	ToM Picture Stories task	.43 – .54	.65 – .81	2	2	2
Strange Stories Task	Faux pas recognition test	.11; not reported	.14; not reported	0 (1)	0 (1)	2
	Reading the Mind in the Eyes Test	.14 – .42; not reported	.21 – .62; not reported	2 (1)	3 (1)	4
	ToM Picture Stories task	.42	.63	1	1	1
ToM Picture Stories task	Faux pas recognition test	.18	.24	0	1	1
	Hinting task	.15	.25	0	1	1
	Reading the Mind in the Eyes Test	.43 – .54	.65 – .81	2	2	2
	Strange Stories Task	.42	.63	1	1	1

study. Other popular tasks have been used for testing older children, such as the Strange Stories task, Animations task, and Hinting task. Some of these tasks show ceiling effects in adults, while others did not (refer to Table 4). It is worth noting that RMET has a child version with fewer items and simpler vocabulary, specifically designed for testing children. Tasks like RMET and Hinting task can be useful for studying how ToM abilities develop from childhood to adulthood and have the potential to provide insight into the continuity of ToM across lifespan. In summary, some tasks originally designed for older children show promise as measures of individual differences in adults. However, like the tasks designed for adults it is unclear what these tasks measure beyond variation in "ToM".

4.4. A programme for future work

The current literature provides considerable prima facie evidence of individual differences in ToM in adults, but much more limited evidence that these differences are psychometrically robust, surprisingly little insight into what this variation might mean, and little evidence that ToM matters for social outcomes in neurotypical adults. New conceptual work and conceptually-motivated empirical work is necessary to clarify in what sense people vary in ToM abilities after they pass the standard assessments of mental state concepts that have been devised for children (e.g., the concepts of desire or belief). Likewise, conceptually-motivated work is necessary to develop a taxonomy of potential ToM components and support the selection of tasks that target such components (Apperly, 2010; Happé et al., 2017; Schaafsma et al., 2015; Warnell and Redcay, 2019). This is likely to require the development of new tasks as well as the systematic examination of existing tasks. In both cases it is essential that the field move towards consistent reporting of information for establishing reliability and validity of measurement. If tasks require component abilities, then examining convergent and discriminant validity is critical to test whether this is reflected in individual differences in performance. The most powerful way to do this is to collect data from multiple tasks in the same participants and test theoretically motivated models of the co-variance. Empirical support for sub-components of ToM would come from meeting two conditions. First, tasks targeting each sub-component should load onto distinct latent variables (demonstrating convergence between tasks testing that sub-component, and divergence from tasks testing other sub-components); second, latent variables for sub-components should nonetheless be correlated (Devine, 2021). Meeting this second condition supplies empirical grounds for saying that the latent variables measure sub-components of a common underlying construct (i.e., ToM). Such a pattern would be similar to findings reported in the executive function literature, which shows shared variance across latent variables that tap on different subdomains, including inhibition, shifting and updating (e.g. Friedman and Miyake, 2017; Miyake et al., 2000). Mapping out the taxonomy of sub-components will help to elucidate the nature of individual differences in adults' ToM.

Finally, it is clearly important to establish that such variance in adults matters for relevant outcomes in real social behaviour, mental health, or wellbeing as much as it appears to matter in childhood (e.g., Hughes and Devine, 2015). The current literature provides a considerable amount of evidence of known-group validity – demonstrating that neurotypical adults perform at higher levels on a given ToM task than a clinical group that is known to have social difficulties. This is clearly of considerable value and interest, but it does not demonstrate that variation in ToM matters for people who do not have a clinical diagnosis. Such evidence is almost entirely lacking at present, and so testing this criterion validity for individual differences in ToM in adults is a clear priority for future work.

4.5. Implications

This review can be used as a reference tool for researchers from all disciplines in psychology who want to examine individual differences in ToM in neurotypical adults to select appropriate task(s). We also suggest a list of attributes concerning reliability and validity that researchers should report when they adopt any of the measures to facilitate future systematic review work in the field, or even meta-analyses. Moreover, the investigation on interrelations among tasks informs us of the potentially multifaceted domain structure of ToM.

4.6. Limitations

One limitation is that we only included English papers for the current review, which may have excluded relevant studies published in other languages. Another limitation is that many measures reviewed lacked comprehensive report of psychometric properties, which limits the confidence of our synthesised results, as it is important to note that lack of evidence is not evidence of absence. Moreover, the current review does not delve into the contentious topic of operationalisation of ToM. We included all measures that purported to be assessing ToM, because our primary objective was to inspect the psychometric properties of such measures. Furthermore, we did not review task durations; measures with good psychometric properties may not be suitable for certain research contexts where time allowed for data collection is limited. Another limitation is that we did not evaluate the relevance of tasks identified to the participants. For example, based on the limited available evidence the MASC shows satisfactory psychometric properties and does not show ceiling effects. However, the video stimuli involve a dinner-date scenario between three white, apparently middle-class Germans aged around thirty to forty. For people who do not speak German it is commonly dubbed into English. While the demographic specificity may help with the realism of the scenario, it also raises the realistic possibility that participants' understanding of the scenario will vary depending upon their own demographics, that is, the task may not demonstrate measurement invariance. This serves to illustrate the general point that it cannot be assumed that the psychometric properties of a test are fixed across contexts. Instead, measurement invariance needs to be established in diverse settings (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014; Nunnally, 1978).

4.7. Future directions

Our findings show that further research on psychometric properties of ToM measures is necessary. We suggest two ways for relevant investigation in the future: the first way is to conduct further research on examining and improving current measures, and the second way is to design new measures that exhibit better psychometric properties.

Recommendations for new research with existing measures. We recommend that measures that exhibit ceiling effects in children should not be used for testing adults. Researchers should always check for ceiling effects. We suggest that more studies that focus on examining psychometric criteria of existing measures be done, and studies adopting such measures should report evidence on reliability and validity. When measures with less satisfactory reliability are adopted, we suggest the use of multiple measures with latent variable modelling to better partial out measurement errors. By using latent variable modelling, the relationships among measures can also be evaluated.

Recommendations for the development of new measures. New measures should aim to achieve good reliability and validity. It is also important to ensure that the measures are relevant and suitable for the participants of interest; age range and culture of participants should be taken into consideration.

4.8. Conclusion

The current review highlights a large evidence gap, whereby the great majority of studies that have examined individual differences in

ToM have not examined whether the tasks are either reliable or valid. In some cases, this is problematic, such as where ceiling effects preclude any meaningful conclusions. The picture emerging from existing evidence provides only very limited confidence in the measurement properties of existing measures, highlighting the need to gain further evidence of reliability and validity of existing measures and to consider development of new measures. Interrelations among measures were inconsistent, which could be due to measurement problems, or due to tasks measuring different aspects of ToM. This highlights the need for empirical work to be aligned with theoretical work on the origins and structure of individual differences in ToM in adults, which should inform both the development of new tasks, and more precise hypotheses about the relevance of ToM for social abilities, mental health and wellbeing.

Author note

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. This research was preregistered on Open Science Framework (OSF).

Data Availability

The data have been uploaded to Open Science Framework, and the link is provided in the manuscript.

Appendix

Summary table of availability of information on scoring attribute and psychometric properties of the top 8 tasks.

Note that the availability of information does not imply the strength of evidence for good psychometric properties. Readers are strongly suggested to refer to the tables and figures in the main text for details.

Measure name	Scoring attribute (Table 1)	Ceiling effect based on POMP score (Table 4) (X = ceiling effect observed)		Types of reliabilities with relevant information (Table 2;Fig. 2) (✓=available)				Types of validities with relevant information (Table 3;Fig. 3) (✓=available)			
		Mean	Range max	C	T	F	R	G	CR	CO	D
Animations Task*	Correctness/propensity			✓			✓	✓		✓	✓
FPRT*	Correctness	X	X	✓		✓	✓	✓	✓	✓	✓
Hinting Task*	Correctness		X	✓	✓	✓	✓	✓	✓	✓	✓
Imposing memory Test	Correctness			✓					✓	✓	✓
MASC	Correctness/(propensity)			✓		✓		✓		✓	✓
RMET	Correctness		X	✓	✓	✓	(✓)	✓	✓	✓	✓
Strange Stories Task*	Correctness	X	X	✓		✓	✓	✓	✓	✓	✓
ToM Picture Stories Task	Correctness	X	X	✓				✓		✓	

* Tested in open-ended format in at least one study.

Reliabilities: C = Convergent validity; T = Test-retest reliability; F = Factor structure; R = Interrater reliability.

Validities: G = Known-group validity; CR = Criterion-related validity; CO = Convergent validity; D = Discriminant validity.

References

Abell, F., Happé, F., Frith, U., 2000. Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. Cogn. Dev. 15 (1), 1–16. https://doi.org/10.1016/S0885-2014(00)00014-9.

Abu-Akel, A.M., Wood, S.J., Hansen, P.C., Apperly, I.A., 2015. Perspective-taking abilities in the balance between autism tendencies and psychosis proneness. Proc. R. Soc. B: Biol. Sci. 282 (1808) https://doi.org/10.1098/rspb.2015.0563.

Achim, A.M., Ouellet, R., Roy, M.-A., Jackson, P.L., 2012. Mentalizing in first-episode psychosis. Psychiatry Res. 196 (2–3), 207–213. https://doi.org/10.1016/j.psychres.2011.10.011.

Adolphs, R., Baron-Cohen, S., Tranel, D., 2002. Impaired recognition of social emotions following amygdala damage. J. Cogn. Neurosci. 14 (8), 1264–1274. https://doi.org/10.1162/089892902760807258.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014. The Standards for Educational and Psychological Testing. American Educational Research Association.

Anastasi, A., 1948. The nature of psychological "traits. Psychol. Rev. 55 (3), 127–138. https://doi.org/10.1037/h0063619.

Apperly, I.A., 2010. Mindreaders: the cognitive basis of "theory of mind.". Psychology Press.

Apperly, I.A., 2012. What is "theory of mind"? Concepts, cognitive processes and individual differences. Q. J. Exp. Psychol. 65 (5), 825–839. https://doi.org/10.1080/17470218.2012.676055.

Apperly, I.A., 2021. Cognitive basis of mindreading in middle childhood and adolescence. Theory Mind Middle Child. Adolesc.: Integr. Mult. Perspect. Routledge, pp. 37–54. https://doi.org/10.4324/9780429326899-4.

Apperly, I.A., Wang, J.J., 2021. Mindreading in adults: Cognitive basis, motivation, and individual differences. In: Ferguson, H.J., Bradford, E.E.F. (Eds.), The Cognitive Basis of Social Interaction Across the Lifespan. Oxford University Press, pp. 96–116. https://doi.org/10.1093/oso/9780198843290.003.0005.

Apperly, I.A., Samson, D., Chiavarino, C., Humphreys, G.W., 2004. Frontal and temporo-parietal lobe contributions to theory of mind: neuropsychological evidence from a false-belief task with reduced language and executive demands. J. Cogn. Neurosci. 16 (10), 1773–1784. https://doi.org/10.1162/0898929042947928.

Apperly, I.A., Samson, D., Humphreys, G.W., 2009b. Studies of adults can inform accounts of theory of mind development. Dev. Psychol. 45 (1), 190–201. https://doi.org/10.1037/a0014098.

Apperly, I.A., Warren, F., Andrews, B.J., Grant, J., Todd, S., 2011. Developmental continuity in theory of mind: speed and accuracy of belief-desire reasoning in children and adults. Child Dev. 82 (5), 1691–1703. https://doi.org/10.1111/j.1467-8624.2011.01635.x.

Atherton, G., Cross, L., 2022. Reading the mind in cartoon eyes: comparing human versus cartoon emotion recognition in those with high and low levels of autistic traits. Psychol. Rep. 125 (3), 1380–1396. https://doi.org/10.1177/0033294120988135.

Aykan, S., Nalçacı, E., 2018. Assessing theory of mind by humor: the humor comprehension and appreciation test (ToM-HCAT. Front. Psychol. 9 https://doi.org/10.3389/fpsyg.2018.01470.

Baksh, R.A., Abrahams, S., Auyeung, B., MacPherson, S.E., 2018. The Edinburgh Social Cognition Test (ESCoT): examining the effects of age on a new measure of theory of mind and social norm understanding. PLoS ONE 13 (4), 1–16. https://doi.org/10.1371/journal.pone.0195818.

Ballespí, S., Vives, J., Sharp, C., Tobar, A., Barrantes-Vidal, N., 2019. Hypermentalizing in social anxiety: evidence for a context-dependent relationship. Front. Psychol. *10*. https://doi.org/10.3389/fpsyg.2019.01501.

Banerjee, R., Watling, D., Caputi, M., 2011. Peer relations and the understanding of faux pas: longitudinal evidence for bidirectional associations. Child Dev. 82 (6), 1887–1905. https://doi.org/10.1111/j.1467-8624.2011.01669.x.

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.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., Robertson, M., 1997. Another advanced test of theory of mind: evidence from very high functioning adults with autism or asperger syndrome. J. Child Psychol. Psychiatry 38 (7), 813–822. https://doi.org/10.1111/j.1469-7610.1997.tb01599.x.

Baron-Cohen, S., O'riordan, M., Stone, V., Jones, R., Plaisted, K., 1999. Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism. J. Autism Dev. Disord. 29 (5), 407–418.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., Plumb, I., 2001. The "reading the mind in the eyes" test revised version: a study with normal adults, and adults with

asperger syndrome or high-functioning autism. J. Child Psychol. Psychiatry 42 (2), 241–251. https://doi.org/10.1111/1469-7610.00715.
Beaudoin, C., Leblanc, É., Gagner, C., Beauchamp, M.H., 2020. Systematic review and inventory of theory of mind measures for young children. Front. Psychol. 10 (January) https://doi.org/10.3389/fpsyg.2019.02905.
Blair, R.J.R., 2000. Impaired social response reversal: a case of `acquired sociopathy'. Brain 123 (6), 1122–1141. https://doi.org/10.1093/brain/123.6.1122.
Bosco, F.M., Colle, L., Fazio, S., De, Bono, A., Ruberti, S., Tirassa, M., 2009. Th.o.m.a.s.: an exploratory assessment of Theory of Mind in schizophrenic subjects. Conscious. Cogn. 18 (1), 306–319. https://doi.org/10.1016/j.concog.2008.06.006.
Bradford, E.E.F., Jentzsch, I., Gomez, J.-C., 2015. From self to social cognition: theory of mind mechanisms and their relation to executive functioning. Cognition 138, 21–34. https://doi.org/10.1016/j.cognition.2015.02.001.
Brewer, N., Young, R.L., Barnett, E., 2017. Measuring theory of mind in adults with autism spectrum disorder. J. Autism Dev. Disord. 47 (7), 1927–1941. https://doi.org/10.1007/s10803-017-3080-x.
Brüne, M., 2003. Theory of mind and the role of IQ in chronic disorganized schizophrenia. Schizophr. Res. 60 (1), 57–64. https://doi.org/10.1016/S0920-9964(02)00162-7.
Brüne, M., 2005. Emotion recognition, 'theory of mind,' and social behavior in schizophrenia. Psychiatry Res. 133 (2–3), 135–147. https://doi.org/10.1016/j.psychres.2004.10.007.
Brunet, E., Sarfati, Y., Hardy-Baylé, M.-C., Decety, J., 2000. A PET investigation of the attribution of intentions with a nonverbal task. NeuroImage 11 (2), 157–166. https://doi.org/10.1006/nimg.1999.0525.
Calso, C., Besnard, J., Allain, P., 2019. Frontal lobe functions in normal aging: metacognition, autonomy, and quality of life. Exp. Aging Res. 45 (1), 10–27. https://doi.org/10.1080/0361073X.2018.1560105.
Canty, A.L., Neumann, D.L., Fleming, J., Shum, D.H.K., 2017. Evaluation of a newly developed measure of theory of mind: The virtual assessment of mentalising ability. Neuropsychol. Rehabil. 27 (5), 834–870. https://doi.org/10.1080/09602011.2015.1052820.
Carpenter, J.M., Green, M.C., Vacharkulksemsuk, T., 2016. Beyond perspective-taking: Mind-reading motivation. Motiv. Emot. 40 (3), 358–374. https://doi.org/10.1007/s11031-016-9544-z.
Channon, S., Crawford, S., 2000. The effects of anterior lesions on performance on a story comprehension test: left anterior impairment on a theory of mind-type task. Neuropsychologia 38 (7), 1006–1017. https://doi.org/10.1016/S0028-3932(99)00154-2.
Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6 (4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284.
Cohen, J., 1992. A power primer. Psychol. Bull. 112 (1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155.
Conway, J.R., Coll, M.-P., Cuve, H.C., Koletsi, S., Bronitt, N., Catmur, C., Bird, G., 2020. Understanding how minds vary relates to skill in inferring mental states, personality, and intelligence. J. Exp. Psychol.: Gen. 149 (6), 1032–1047. https://doi.org/10.1037/xge0000704.
Corcoran, R., Mercer, G., Frith, C.D., 1995. Schizophrenia, symptomatology and social inference: investigating "theory of mind" in people with schizophrenia. Schizophr. Res. 17 (1), 5–13. https://doi.org/10.1016/0920-9964(95)00024-G.
Corcoran, R., Cahill, C., Frith, C., 1997. The appreciation of visual jokes in people with schizophrenia: a study of 'mentalizing' ability. Schizophr. Res. 24 (3), 319–327. https://doi.org/10.1016/S0920-9964(96)00117-X.
Cortina, J.M., 1993. What is coefficient alpha? An examination of theory and applications. J. Appl. Psychol. 78 (1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98.
Derksen, D.G., Hunsche, M.C., Giroux, M.E., Connolly, D.A., Bernstein, D.M., 2018. A systematic review of theory of mind's precursors and functions. Z. Fur Psychol. / J. Psychol. 226 (2), 87–97. https://doi.org/10.1027/2151-2604/a000325.
Devine, R.T., 2021. Individual differences in theory of mind in middle childhood and adolescence. Theory Mind Middle Child. Adolesc.: Integr. Mult. Perspect. Routledge/Taylor & Francis Group, pp. 55–76. https://doi.org/10.4324/9780429326899-5.
Devine, R.T., Apperly, I.A., 2022. Willing and able? Theory of mind, social motivation, and social competence in middle childhood and early adolescence. Dev. Sci. 25 (1), 1–14. 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 Dev. 84 (3), 989–1003. https://doi.org/10.1111/cdev.12017.
Devine, R.T., Hughes, C., 2016. Measuring theory of mind across middle childhood: reliability and validity of the Silent Films and Strange Stories tasks. J. *Exp. Child Psychol.* 149, 23–40. https://doi.org/10.1016/j.jecp.2015.07.011.
Devine, R.T., Kovatchev, V., Traynor, I.G., Smith, P., Lee, M., 2023. Machine learning and deep learning systems for automated measurement of "advanced" theory of mind: reliability and validity in children and adolescents. Psychol. Assess. 35 (2), 165–177. https://doi.org/10.1037/pas0001186.
Dewey, M., 1991. Living with Asperger's syndrome. Autism and Asperger Syndrome. Cambridge University Press,, pp. 184–206. https://doi.org/10.1017/CBO9780511526770.006.
Dodell-Feder, D., Lincoln, S.H., Coulson, J.P., Hooker, C.I., 2013. Using fiction to assess mental state understanding: a new task for assessing theory of mind in adults. PLoS ONE 8 (11), 1–14. https://doi.org/10.1371/journal.pone.0081279.
Dodich, A., Cerami, C., Canessa, N., Crespi, C., Iannaccone, S., Marcone, A., Realmuto, S., Lettieri, G., Perani, D., Cappa, S.F., 2015. A novel task assessing intention and emotion attribution: Italian standardization and normative data of the Story-based Empathy Task. Neurol. Sci. 36 (10), 1907–1912. https://doi.org/10.1007/s10072-015-2281-3.
Dwyer, K., David, A.S., McCarthy, R., McKenna, P., Peters, E., 2020. Linguistic alignment and theory of mind impairments in schizophrenia patients' dialogic interactions. Psychol. Med. 50 (13), 2194–2202. https://doi.org/10.1017/S0033291719002289.
Dyck, M.J., Ferguson, K., Shochet, I.M., 2001. Do autism spectrum disorders differ from each other and from non-spectrum disorders on emotion recognition tests? Eur. Child Adolesc. Psychiatry 10 (2), 105–116. https://doi.org/10.1007/s007870170033.
Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J.K., Wolf, O.T., Convit, A., 2006. Introducing MASC: a movie for the assessment of social cognition. J. Autism Dev. Disord. 36 (5), 623–636. https://doi.org/10.1007/s10803-006-0107-0.
Dziobek, I., Rogers, K., Fleck, S., Bahnemann, M., Heekeren, H.R., Wolf, O.T., Convit, A., 2008. Dissociation of cognitive and emotional empathy in adults with asperger syndrome using the multifaceted empathy test (MET. J. Autism Dev. Disord. 38 (3), 464–473. https://doi.org/10.1007/s10803-007-0486-x.
El Haj, M., Antoine, P., Nandrino, J.L., 2017. When deception influences memory: the implication of theory of mind. Q. J. Exp. Psychol. 70 (7), 1166–1173. https://doi.org/10.1080/17470218.2016.1173079.
Fleiss, J.L., 1986. The design and analysis of clinical experiments. Wiley.
Friedman, N.P., Miyake, A., 2017. Unity and diversity of executive functions: Individual differences as a window on cognitive structure. Cortex 86, 186–204. https://doi.org/10.1016/j.cortex.2016.04.023.
Frith, C.D., Corcoran, R., 1996. Exploring 'theory of mind' in people with schizophrenia. Psychol. Med. 26 (3), 521–530. https://doi.org/10.1017/S0033291700035601.
Fu, I., Chen, K., Liu, M., Jiang, D., Hsieh, C.-L., Lee, S.-C., 2023. A systematic review of measures of theory of mind for children. Dev. Rev. 67 (1), 101061 https://doi.org/10.1016/j.dr.2022.101061.
Gallagher, H., Happé, F., Brunswick, N., Fletcher, P., Frith, U., Frith, C., 2000. Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. Neuropsychologia 38 (1), 11–21. https://doi.org/10.1016/S0028-3932(99)00053-6.
Gallant, C., Good, D., 2020. Examining the "reading the mind in the eyes test" as an assessment of subtle differences in affective theory of mind after concussion. Clin. Neuropsychol. 34 (2), 296–317. https://doi.org/10.1080/13854046.2019.1612946.
German, T.P., Hehman, J.A., 2006. Representational and executive selection resources in "theory of mind": evidence from compromised belief-desire reasoning in old age. Cognition 101 (1), 129–152. https://doi.org/10.1016/j.cognition.2005.05.007.
Gignac, G.E., Szodorai, E.T., 2016. Effect size guidelines for individual differences researchers. Personal. Individ. Differ. 102, 74–78. https://doi.org/10.1016/j.paid.2016.06.069.
Gilpin, A.R., 1993. Table for conversion of Kendall's Tau to Spearman's Rho within the context of measures of magnitude of effect for meta-analysis. Educ. *Psychol. Meas.* 53 (1), 87–92. https://doi.org/10.1177/0013164493053001007.
Girardi, A., MacPherson, S.E., Abrahams, S., 2011. Deficits in emotional and social cognition in amyotrophic lateral sclerosis. Neuropsychology 25 (1), 53–65. https://doi.org/10.1037/a0020357.
Golan, O., Baron-Cohen, S., Hill, J., 2006. The cambridge mindreading (CAM) face-voice battery: testing complex emotion recognition in adults with and without asperger syndrome. J. Autism Dev. Disord. 36 (2), 169–183. https://doi.org/10.1007/s10803-005-0057-y.
Golan, O., Baron-Cohen, S., Hill, J.J., Golan, Y., 2006. The "Reading the Mind in Films" Task: Complex emotion recognition in adults with and without autism spectrum conditions. Soc. Neurosci. 1 (2), 111–123. https://doi.org/10.1080/17470910600980986.
Golan, O., Baron-Cohen, S., Hill, J.J., Rutherford, M.D., 2007. The 'reading the mind in the voice' test-revised: a study of complex emotion recognition in adults with and without autism spectrum conditions. J. Autism Dev. Disord. 37 (6), 1096–1106. https://doi.org/10.1007/s10803-006-0252-5.
Gopnik, A., Astington, J.W., 1988. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. Child Dev. 59 (1), 26–37. https://doi.org/10.1111/j.1467-8624.1988.tb03192.x.
Happe, F., 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. J. Autism Dev. Disord. 24 (2) https://doi.org/10.1007/BF02172093.
Happe, F., Frith, U., 1996. The neuropsychology of autism. Brain 119 (4), 1377–1400. https://doi.org/10.1093/brain/119.4.1377.
Happé, F., Brownell, H., Winner, E., 1999. Acquired `theory of mind' impairments following stroke. Cognition 70 (3), 211–240. https://doi.org/10.1016/S0010-0277(99)00005-0.
Happé, F., Cook, J.L., Bird, G., 2017. The structure of social cognition: in(ter)dependence of sociocognitive processes. Annu. Rev. Psychol. 68 (1), 243–267. https://doi.org/10.1146/annurev-psych-010416-044046.
Hedge, C., Powell, G., Sumner, P., 2018. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Methods 50 (3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1.
Henry, A., Tourbah, A., Chaunu, M.-P., Rumbach, L., Montreuil, M., Bakchine, S., 2011. Social Cognition Impairments in Relapsing-Remitting Multiple Sclerosis. J. Int. Neuropsychol. Soc. 17 (6), 1122–1131. https://doi.org/10.1017/S1355617711001147.
Hughes, C., 2016. Theory of mind grows up: Reflections on new research on theory of mind in middle childhood and adolescence. J. Exp. Child Psychol. 149, 1–5. https://doi.org/10.1016/j.jecp.2016.01.017.

Hughes, C., Devine, R.T., 2015. Individual differences in theory of mind from preschool to adolescence: achievements and directions. Child Dev. Perspect. 9 (3), 149–153. https://doi.org/10.1111/cdep.12124.
Hughes, C., Adlam, A., Happé, F., Jackson, J., Taylor, A., Caspi, A., 2000. Good test-retest reliability for standard and advanced false-belief tasks across a wide range of abilities. J. Child Psychol. Psychiatry Allied Discip. 41 (4), 483–490. https://doi.org/10.1017/S0021963099005533.
Ickes, W., Buysse, A., Pham, H., Rivers, K., Erickson, J.R., Hancock, M., Kelleher, J., Gesn, P.R., 2000. On the difficulty of distinguishing "good" and "poor" perceivers: a social relations analysis of empathic accuracy data. Pers. Relatsh. 7 (2), 219–234. https://doi.org/10.1111/j.1475-6811.2000.tb00013.x.
Imuta, K., Henry, J.D., Slaughter, V., Selcuk, B., Ruffman, T., 2016. Theory of mind and prosocial behavior in childhood: a meta-analytic review. Dev. Psychol. 52 (8), 1192–1205. https://doi.org/10.1037/dev0000140.
Kanske, P., Böckler, A., Trautwein, F.M., Singer, T., 2015. Dissecting the social brain: Introducing the EmpaToM to reveal distinct neural networks and brain-behavior relations for empathy and Theory of Mind. NeuroImage 122, 6–19. https://doi.org/10.1016/j.neuroimage.2015.07.082.
Karmakar, A., Dogra, A.K., 2019. Assessment of theory of mind in adults: beyond false belief tasks. Act. Nerv. Super. 61 (3), 142–146. https://doi.org/10.1007/s41470-019-00028-1.
Keysar, B., Barr, D.J., Balin, J.A., Brauner, J.S., 2000. Taking perspective in conversation: the role of mutual knowledge in comprehension. Psychol. Sci. 11 (1), 32–38. https://doi.org/10.1111/1467-9280.00211.
Keysar, B., Lin, S., Barr, D.J., 2003. Limits on theory of mind use in adults. Cognition 89, 25–41. https://doi.org/10.1016/S0010-0277(03)00064-7.
Kinderman, P., Dunbar, R., Bentall, R.P., 1998. Theory-of-mind deficits and causal attributions. Br. J. Psychol. 89 (2), 191–204. https://doi.org/10.1111/j.2044-8295.1998.tb02680.x.
Klin, A., 2000. Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and asperger syndrome: the social attribution task. J. Child Psychol. Psychiatry 41 (7), 831–846. https://doi.org/10.1111/1469-7610.00671.
Kosmidis, M.H., Giannakou, M., Garyfallos, G., Kiosseoglou, G., Bozikas, V.P., 2011. The impact of impaired "theory of mind" on social interactions in schizophrenia. J. Int. Neuropsychol. Soc. 17 (3), 511–521. https://doi.org/10.1017/S1355617711000300.
Koster-Hale, J., Dodell-Feder, D., & Saxe, R. (2012). [unpublished instrument].
Langdon, R., Coltheart, M., Ward, P.B., Catts, S.V., 2002. Disturbed communication in schizophrenia: the role of poor pragmatics and poor mind-reading. Psychol. Med. 32 (7), 1273–1284. https://doi.org/10.1017/S0033291702006396.
Langdon, Robyn, Michie, P.T., Ward, P.B., McConaghy, N., Catts, S.V., Coltheart, M., 1997. Defective self and/or other mentalising in schizophrenia: a cognitive neuropsychological approach. Cogn. Neuropsychiatry 2 (3), 167–193. https://doi.org/10.1080/135468097396324.
Lawson, J., Baron-Cohen, S., Wheelwright, S., 2004. Empathising and systemising in adults with and without asperger syndrome. J. Autism Dev. Disord. 34 (3), 301–310. https://doi.org/10.1023/B:JADD.0000029552.42724.1b.
Licata, M., Zietlow, A.-L., Träuble, B., Sodian, B., Reck, C., 2016. Maternal emotional availability and its association with maternal psychopathology, attachment style insecurity and theory of mind. Psychopathology 49 (5), 334–340. https://doi.org/10.1159/000447781.
MacCann, C., Roberts, R.D., 2008. New paradigms for assessing emotional intelligence: Theory and data. Emotion 8 (4), 540–551. https://doi.org/10.1037/a0012746.
Mahy, C.E.V., Moses, L.J., Pfeifer, J.H., 2014. How and where: theory-of-mind in the brain. Dev. Cogn. Neurosci. 9, 68–81. https://doi.org/10.1016/j.dcn.2014.01.002.
Matheson, G.J., 2019. We need to talk about reliability: making better use of test-retest studies for study design and interpretation. PeerJ 7 (5), e6918. https://doi.org/10.7717/peerj.6918.
Mayes, L.C., Klin, A., Tercyak, K.P., Cicchetti, D.V., Cohen, D.J., 1996. Test-retest reliability for false-belief tasks. J. Child Psychol. Psychiatry 37 (3), 313–319. https://doi.org/10.1111/j.1469-7610.1996.tb01408.x.
McDonald, S., Flanagan, S., Rollins, J., Kinch, J., 2003. TASIT: a new clinical tool for assessing social perception after traumatic brain injury. J. Head. Trauma Rehabil. 18 (3), 219–238. https://doi.org/10.1097/00001199-200305000-00001.
McGarry, K.A., West, M., Hogan, K.F., 2021. Perspective-taking and social competence in adults. Adv. Cogn. Psychol. 17 (2), 129–135. https://doi.org/10.5709/ACP-0323-5.
Mendoza, J.L., Mumford, M., 1987. Corrections for Attenuation and Range Restriction on the Predictor. J. Educ. Stat. 12 (3), 282. https://doi.org/10.2307/1164688.
Miyake, A., Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A., Wager, T.D., 2000. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. Cogn. Psychol. 41 (1), 49–100. https://doi.org/10.1006/cogp.1999.0734.
Mordal, J., Gundersen, Ø., Bramness, J.G., 2010. Norwegian version of the Mini-International Neuropsychiatric Interview: feasibility, acceptability and test-retest reliability in an acute psychiatric ward. Eur. Psychiatry 25 (3), 172–177. https://doi.org/10.1016/j.eurpsy.2009.02.004.
Murphy, B.A., Lilienfeld, S.O., 2019. Are self-report cognitive empathy ratings valid proxies for cognitive empathy ability? Negligible meta-analytic relations with behavioral task performance. Psychol. Assess. 31 (8), 1062–1072. https://doi.org/10.1037/pas0000732.
Murray, K., Johnston, K., Cunnane, H., Kerr, C., Spain, D., Gillan, N., Hammond, N., Murphy, D., Happé, F., 2017. A new test of advanced theory of mind: The "Strange Stories Film Task" captures social processing differences in adults with autism spectrum disorders. Autism Res. 10 (6), 1120–1132. https://doi.org/10.1002/aur.1744.
Nilsen, E.S., Duong, D., 2013. Depressive symptoms and use of perspective taking within a communicative context. Cogn. Emot. 27 (2), 335–344. https://doi.org/10.1080/02699931.2012.708648.
Nunnally, J.C., 1978. Psychometric Theory, 2nd ed...,. McGraw-Hill,.
Oakley, B.F.M., Brewer, R., Bird, G., Catmur, C., 2016. Theory of mind is not theory of emotion: a cautionary note on the Reading the Mind in the Eyes Test. J. Abnorm. Psychol. 125 (6), 818–823. https://doi.org/10.1037/abn0000182.
Orbelo, D.M., Grim, M.A., Talbott, R.E., Ross, E.D., 2005. Impaired comprehension of affective prosody in elderly subjects is not predicted by age-related hearing loss or age-related cognitive decline. J. Geriatr. Psychiatry Neurol. 18 (1), 25–32. https://doi.org/10.1177/0891988704272214.
Osterhaus, C., Bosacki, S.L., 2022. Looking for the lighthouse: a systematic review of advanced theory-of-mind tests beyond preschool. Dev. Rev. 64 (February), 101021 https://doi.org/10.1016/j.dr.2022.101021.
Ouellet, J., Scherzer, P.B., Rouleau, I., Métras, P., Bertrand-Gauvin, C., Djerroud, N., Boisseau, É., Duquette, P., 2010. Assessment of social cognition in patients with multiple sclerosis. J. Int. Neuropsychol. Soc. 16 (2), 287–296. https://doi.org/10.1017/S1355617709991329.
Paal, T., Bereczkei, T., 2007. Adult theory of mind, cooperation, Machiavellianism: the effect of mindreading on social relations. Personal. Individ. Differ. 43 (3), 541–551. https://doi.org/10.1016/j.paid.2006.12.021.
Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., Moher, D., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372, n71. https://doi.org/10.1136/bmj.n71.
Perner, J., Wimmer, H., 1985. "John thinks that Mary thinks that…" attribution of second-order beliefs by 5- to 10-year-old children. J. Exp. Child Psychol. 39 (3), 437–471. https://doi.org/10.1016/0022-0965(85)90051-7.
Perner, J., Leekam, S.R., Wimmer, H., 1987. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. Br. J. Dev. Psychol. 5 (2), 125–137. https://doi.org/10.1111/j.2044-835x.1987.tb01048.x.
Petersen, I.T., Hoyniak, C.P., McQuillan, M.E., Bates, J.E., Staples, A.D., 2016. Measuring the development of inhibitory control: The challenge of heterotypic continuity. Developmental Review 40 (3), 25–71. https://doi.org/10.1016/j.dr.2016.02.001.
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 Dev. 83 (2), 469–485. https://doi.org/10.1111/j.1467-8624.2011.01728.x.
Pickup, G.J., Frith, C.D., 2001. Theory of mind impairments in schizophrenia: symptomatology, severity and specificity. Psychol. Med. 31 (2), 207–220. https://doi.org/10.1017/S0033291701003385.
Pisani, S., Murphy, J., Conway, J., Millgate, E., Catmur, C., Bird, G., 2021. The relationship between alexithymia and theory of mind: a systematic review. Neurosci. Biobehav. Rev. 131, 497–524. https://doi.org/10.1016/j.neubiorev.2021.09.036.
Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind. Behav. Brain Sci. 1 (4), 515–526. https://doi.org/10.1017/S0140525X00076512.
Quesque, F., Rossetti, Y., 2020. What do theory-of-mind tasks actually measure? Theory and practice. Perspect. Psychol. Sci. 15 (2), 384–396. https://doi.org/10.1177/1745691619896607.
Qureshi, A.W., Monk, R.L., Samson, D., Apperly, I.A., 2020. Does interference between self and other perspectives in theory of mind tasks reflect a common underlying process? Evidence from individual differences in theory of mind and inhibitory control. Psychon. Bull. Rev. 27 (1), 178–190. https://doi.org/10.3758/s13423-019-01656-z.
Rakoczy, H., Wandt, R., Thomas, S., Nowak, J., Kunzmann, U., 2018. Theory of mind and wisdom: the development of different forms of perspective-taking in late adulthood. Br. J. Psychol. 109 (1), 6–24. https://doi.org/10.1111/bjop.12246.
Revelle, W., Condon, D.M., 2019. Reliability from α to ω: a tutorial. Psychol. Assess. 31 (12), 1395–1411. https://doi.org/10.1037/pas0000754.
Rice, K., Redcay, E., 2015. Spontaneous mentalizing captures variability in the cortical thickness of social brain regions. Soc. Cogn. Affect. Neurosci. 10 (3), 327–334. https://doi.org/10.1093/scan/nsu081.
Roberts, D.L., Fiszdon, J., Tek, C., 2011. Ecological validity of the social cognition screening questionnaire (SCSQ). *Abstr. 13th Int. Congr. Schizophr. Res. (ICOSR)* 280. https://doi.org/10.1093/schbul/sbq173.
Rönkkö, M., Cho, E., 2022. An updated guideline for assessing discriminant validity. Organ. Res. Methods 25 (1), 6–14. https://doi.org/10.1177/1094428120968614.
Rosenblau, G., Kliemann, D., Heekeren, H.R., Dziobek, I., 2015. Approximating implicit and explicit mentalizing with two naturalistic video-based tasks in typical development and autism spectrum disorder. J. Autism Dev. Disord. 45 (4), 953–965. https://doi.org/10.1007/s10803-014-2249-9.
Rust, J., Kosinski, M., Stillwell, D., 2020. Modern Psychometrics, *fourth ed.* Routledge. https://doi.org/10.4324/9781315637686.
Rutherford, M.D., 2004. The effect of social role on theory of mind reasoning. Br. J. Psychol. 95 (1), 91–103. https://doi.org/10.1348/000712604322779488.
Samson, D., Apperly, I.A., Braithwaite, J.J., Andrews, B.J., Bodley Scott, S.E., 2010. Seeing it their way: evidence for rapid and involuntary computation of what other people see. J. Exp. Psychol.: Hum. Percept. Perform. 36 (5), 1255–1266. https://doi.org/10.1037/a0018729.
Sarfati, Y., Hardybayle, M., Besche, C., Widlocher, D., 1997. Attribution of intentions to others in people with schizophrenia: a non-verbal exploration with comic strips. Schizophr. Res. 25 (3), 199–209. https://doi.org/10.1016/S0920-9964(97)00025-X.
Schaafsma, S.M., Pfaff, D.W., Spunt, R.P., Adolphs, R., 2015. Deconstructing and reconstructing theory of mind. Trends Cogn. Sci. 19 (2), 65–72. https://doi.org/10.1016/j.tics.2014.11.007.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J., 2014. Fractionating theory of mind: A meta-analysis of functional brain imaging studies. Neurosci. Biobehav. Rev. 42, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009.

Sebastian, C.L., Fontaine, N.M.G., Bird, G., Blakemore, S.-J., De Brito, S.A., McCrory, E.J. P., Viding, E., 2012. Neural processing associated with cognitive and affective theory of Mind in adolescents and adults. Soc. Cogn. Affect. Neurosci. 7 (1), 53–63.

Shah, P., Catmur, C., Bird, G., 2017. From heart to mind: linking interoception, emotion, and theory of mind. Cortex 93, 220–223. https://doi.org/10.1016/j.cortex.2017.02.010.

Shamay-Tsoory, S.G., Shur, S., Barcai-Goodman, L., Medlovich, S., Harari, H., Levkovitz, Y., 2007. Dissociation of cognitive from affective components of theory of mind in schizophrenia. Psychiatry Res. 149 (1–3), 11–23. https://doi.org/10.1016/j.psychres.2005.10.018.

Shaw, P., Lawrence, E.J., Radbourne, C., Bramham, J., Polkey, C.E., David, A.S., 2004. The impact of early and late damage to the human amygdala on 'theory of mind' reasoning. Brain 127 (7), 1535–1548. https://doi.org/10.1093/brain/awh168.

Sommerville, J.A., Bernstein, D.M., Meltzoff, A.N., 2013. Measuring beliefs in centimeters: private knowledge biases preschoolers' and adults' representation of others' beliefs. Child Dev. 84 (6), 1846–1854. https://doi.org/10.1111/cdev.12110.

Sullivan, S., Ruffman, T., 2004. Social understanding: how does it fare with advancing years. Br. J. Psychol. 95 (1), 1–18. https://doi.org/10.1348/000712604322779424.

Tahazadeh, S., Barahmand, U., Yaghooti, F., Nazari, M.A., 2020. Mind reading in films task to assess social cognitive deficits in autism spectrum conditions. J. Evid. -Based Psychother. 20 (2), 79–100. https://doi.org/10.24193/jebp.2020.2.13.

Wang, X., Su, Y., Pei, M., Hong, M., 2021. How self-other control determines individual differences in adolescents' theory of mind. Cogn. Dev. 57, 101007 https://doi.org/10.1016/j.cogdev.2021.101007.

Wang, Z., Devine, R.T., Wong, K.K., Hughes, C., 2016. Theory of mind and executive function during middle childhood across cultures. J. Exp. Child Psychol. 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, 103997. https://doi.org/10.1016/j.cognition.2019.06.009.

Watson, A.C., Nixon, C.L., Wilson, A., Capage, L., 1999. Social interaction skills and theory of mind in young children. Dev. Psychol. 35 (2), 386–391. https://doi.org/10.1037/0012-1649.35.2.386.

Weimer, A.A., Warnell, K.R., Ettekal, I., Cartwright, K.B., Guajardo, N.R., Liew, J., 2021. Correlates and antecedents of theory of mind development during middle childhood and adolescence: an integrated model. Dev. Rev. 59 (December 2020), 100945 https://doi.org/10.1016/j.dr.2020.100945.

Weinstein, N.Y., Whitmore, L.B., Mills, K.L., 2022. Individual differences in mentalizing tendencies. Collabra: Psychol. 8 (1), 1–22. https://doi.org/10.1525/collabra.37602.

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

Yirmiya, N., Erel, O., Shaked, M., Solomonica-Levi, D., 1998. Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. Psychol. Bull. 124 (3), 283–307. https://doi.org/10.1037/0033-2909.124.3.283.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. Proceedings of the National Academy of Sciences, *104*(20), 8235–8240. https://doi.org/10.1073/pnas.0701408104.

Ziatabar Ahmadi, S.Z., Jalaie, S., Ashayeri, H., 2015. Validity and reliability of published comprehensive theory of mind tests for normal preschool children: a systematic. Iran. J. Psychiatry 10 (4), 214–224.