Development and Validation of the Four-Item Mentalising Index (FIMI)

Rachel A. Clutterbuck, University of Bath

Mitchell J. Callan, University of Bath

Emily C. Taylor, University of Bath

Lucy A. Livingston, Cardiff University

Punit Shah, University of Bath

Author Note

Rachel A. Clutterbuck, Mitchell J. Callan, Emily C. Taylor, & Punit Shah, Department of Psychology, University of Bath, Bath, U.K.

Lucy A. Livingston, School of Psychology, Cardiff University, Cardiff, U.K. and Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, U.K.

RAC, LAL, and PS conceived the instrument, RAC, LAL, MJC, and PS collected and analysed the data, and all authors drafted, read, and approved the final version of the manuscript. The authors thank Zoe Freeman, Laura Smith, and Lois Player for acting as independent raters and/or assistance with data collection. RAC and ECT are supported by doctoral studentships from the Economic and Social Research Council and the Whorrod Foundation, respectively. All authors declare no conflict of interest.

Correspondence concerning this article should be addressed to p.shah@bath.ac.uk. Department of Psychology, University of Bath, Bath, BA2 7AY, U.K.

Abstract

Mentalising, also known as 'Theory of Mind', is the ability to understand and infer the cognitions of others, such as their perceptions, intentions, and beliefs. Although several tools have been designed to measure mentalising in adults, there exist methodological and practical limitations. Many of the existing measures conflate mentalising with similar constructs (e.g., empathy), and most are lengthy measures that are unsuitable for large population-based studies and clinical practice. These issues are currently hampering clinical and non-clinical investigations into mentalising and related social-cognitive abilities. Drawing on questionnaire measures of social cognition, we conceived a self-report mentalising scale, the Four-Item Mentalising Index (FIMI; Studies 1a & b). The FIMI was developed through a series of studies examining its factor structure and reliability (Studies 2a & b) and by testing its construct validity against a cognitive mentalising task, autistic traits, and comparing scores in autistic and non-autistic people (Studies 3a & b). Together, we demonstrate that the FIMI is a conceptually and methodologically robust tool for measuring mentalising ability in the general population, including autistic and non-autistic people. Future research directions and practical (clinical) applications of the scale are discussed, with a focus on improving understanding and management of (a)typical mentalising ability.

Keywords: Autism; Cognition; Empathy; Mentalising; Self-report; Theory of Mind

Public Significance Statement: We developed a reliable questionnaire to measure how well people understand what others are thinking. It was validated against an objective test and by giving the questionnaire to autistic people who are known to have difficulties understanding others. We discuss how this tool will be useful in future large-scale studies and clinical practice.

‘Mentalising’ or ‘Theory of Mind’ is the social-cognitive ability to understand and infer the mental states of oneself and others, including beliefs, intentions, and desires (see Happé et al., 2017). A large body of research has examined mentalising in children, but less research has explored mentalising in adulthood (Livingston & Happé, in press). Investigating mentalising in adulthood is critically important to understand social-cognitive changes in our aging population (Henry et al., 2013) and clinical phenomena, such as autism, which are characterised by mentalising difficulties (e.g., Lever & Geurts, 2016).

There has been a growing interest in this field, with several measures being designed to examine mentalising in typical and atypical adults. These include cognitive tasks that involve inferring character’s mental states from vignettes (e.g., Happé, 1994), animations (e.g., White et al., 2011), videos (e.g., Murray et al., 2017), and images of the eye region (e.g., Olderbak et al., 2015). However, practical issues with these measures, such as their complexity and administration time, limit their use in many research and clinical settings, thereby constraining current understanding of (a)typical mentalising across the lifespan. For example, the widely used Frith-Happé Animations Task requires presentation of several videos followed by recording, transcription, and coding of participants’ responses, which is not possible in certain research designs or clinical practice (see Livingston, Carr et al., 2019).

There are also concerns about the validity of existing measures; they often have poor or untested psychometric properties and concerns have been raised about whether they are truly measuring mentalising or other social-cognitive processes such as emotion processing (see Olderbak et al., 2019). Accordingly, recent research has found poor associations between self-report and cognitive measures of social cognition (e.g., empathy; Murphy & Lilienfeld, 2019), as well as relatively low correlations between different mentalising tasks (Morrison et al., 2019; Warnell & Redcay, 2019; Navarro et al., in press). This is likely because these measures are combining and conflating several constructs, including mentalising, emotion perception, and empathy (Quesque & Rossetti, 2020). Nonetheless, by carefully distinguishing between social-cognitive constructs in questionnaires and simplifying their conceptual complexity (see McGrath, 2005), we suggest that it is possible to develop self-report mentalising measures and improve their use in applied settings (e.g., in clinical practice).

In view of these practical, methodological, and conceptual issues in previous research, researchers and clinicians would benefit from the development of a short, valid, and easily administered self-report measure of mentalising. Beyond its practical importance, such a measure also has potential to advance our theoretical understanding of mentalising, particularly informing debates about its (non)overlap with related psychological constructs. To this end, we aimed to develop and validate the Four-Item Mentalising Index (FIMI). Drawing on questionnaire measures of social cognition and using factor analysis, we selected items for the FIMI (Studies 1a & b), followed by examination of its factor structure, measurement invariance by sex, internal consistency, and test-retest reliability (Studies 2a & b). Finally, we examined its validity against a cognitive mentalising task, autistic traits, and comparisons between autistic and non-autistic people (Studies 3a & b).

Study 1: Conceptual Analysis and Item Selection

We first drew upon items from the Questionnaire of Cognitive and Affective Empathy (QCAE; Reniers et al., 2011) to develop the FIMI, which combines several existing empathy scales. It has subscales measuring two types of empathy: *affective* (experiencing others' emotional states) and *cognitive* (understanding others' emotional states). Cognitive empathy and mentalising are sometimes considered interchangeable (e.g., Rueda et al., 2015), but evidence suggests these constructs, with different underlying neurobiological mechanisms (Preckel et al., 2018), are not equivalent. Such distinctions between cognitive empathy and mentalising are critically important for studying different social-cognitive processes, particularly in clinical populations (Olderbak et al., 2019). Indeed, when developing the QCAE, Reniers et al. (2011) noted that, "cognitive empathy is concerned with the attribution of emotions as opposed to cognitions, and as such the two constructs [mentalising and cognitive empathy] are potentially dissociable" (p.85). Accordingly, we propose that mentalising and cognitive empathy are related but separable constructs, whereby we define cognitive empathy as understanding of others' emotions, whereas mentalising is the understanding of *non-emotional* mental states. Following this line of enquiry, a cursory inspection of the QCAE highlighted that some of its items were more closely aligned with the conceptual definition of mentalising than empathy. To confirm this observation, using emotional inference as the critical feature that distinguishes empathy from mentalising, we asked a panel to identify QCAE items that *did not* have emotion-related language and then conducted factor analyses to select mentalising items.

Study 1a. The factorial structure of the QCAE has previously been examined (e.g., Reniers et al., 2011), but not for the purpose of excluding emotional and identifying mentalising items. Given the subtle differences in emotional language which may distinguish mentalising from cognitive empathy, it was unlikely that this distinction could solely be made through a data-driven approach. Therefore, following scale development procedures for interpreting text (Carpenter, 2018), non-emotional QCAE items were manually identified. Four raters—two authors and two independent experts—independently selected QCAE items that were free from emotion-related language. They reliably selected 9 items (Krippendorff's $\alpha = .89$) with minimal discrepancy, which was resolved through discussion until consensus. Items including emotional words (e.g., "*feel*") were excluded, whereas items without emotional language were selected (Table 1).

We then analysed an existing dataset of an online study in which participants (N = 660, 326 females, aged 18-85, $M_{age} = 35.24$, $SD_{age} = 12.02$ years) had completed the QCAE as a measure of empathy (see Shah et al., 2019) using the standard 4-point response scale (1 = *strongly disagree*, 2 = *slightly disagree*, 3 = *slightly agree*, 4 = *strongly agree*). Exploratory Factor Analysis (EFA), with principal axis factoring and Oblimin rotation, examined the structure underlying the 9 non-emotional QCAE items that had manually been identified.[1] This revealed 2 moderately correlated (r = .55) factors with initial SS loadings greater than 1 (2.04 and 1.64; 41% total variance explained). Examination of the individual items (Table 1) suggested that 5 items in Factor 1 pertained to social cognition potentially involving but not directly measuring mentalising (e.g., "I am good at predicting what someone will do"), whereas the 4 items in Factor 2 clearly aligned with the conceptual definition of mentalising (e.g., "I sometimes try to understand my friends better by imagining how things look from their perspective").

Further empirical support for the distinction between emotional and non-emotional QCAE items, as identified by the four raters, was sought using Confirmatory Factor Analysis (CFA). This indicated that a 2-factor model, separating 9 non-emotional QCAE items from the 22 emotional items, showed significantly better fit than a 1-factor solution of 31 items, $\chi^2_{\text{difference}}$ (1) = 29.57, p < .001. Further, a 2-factor model, separating 4 potential mentalising items from the remaining 27 QCAE items, showed significantly better fit than a 1-factor solution, $\chi^2_{\text{difference}}$ (1) = 114.67, p < .001. Together, these results demonstrated that several

[1] Across studies, factor analyses were performed using the Psych (Revelle, 2018, v.1.8.12), GPA rotation (Bernaards & Jennrich, 2005, v.2014.11-1), Lavaan (Rosseel, 2012, v.0.6-3) and/or semTools (Jorgensen et al., 2018, v.0.5-1) packages in R. Models were estimated using the default maximum likelihood estimation in Lavaan. Because the QCAE/FIMI has a 4-point Likert scale, we supplemented all our main CFAs using an estimator for ordinal variables, which yielded the same conclusions (see Supplementary Materials).

QCAE items were measuring a construct that was distinguishable from empathy, thereby supporting further investigation of their suitability for a mentalising scale.

Study 1b. To address concerns with Study 1a – because it drew from an existing dataset of the full QCAE and other measures – we recruited a new sample of participants to complete an online study of just the 9 non-emotional items selected from the QCAE. This sample (N = 669) was recruited via Amazon's Mechanical Turk (Mturk) so that it was similar to Study 1a. English-speaking participants, recruited from the United States (US) and United Kingdom (UK), reported their age and sex, demonstrating that the sample contained a wide age range of males and females (342 females, aged 18-80, M_{age} = 36.39, SD_{age} = 12.14 years). There were no other inclusion or exclusion criteria, and there were no attention checks as the study was short in duration. All studies were approved by the local ethics committee, and participants gave informed consent and were debriefed following their participation. As this study sought to confirm the EFA results in Study 1a, response data were submitted to a CFA. This showed that a 2-factor solution, χ^2 (26) = 95.42, p < .001; CFI = .95; TLI = .93; RMSEA = .06; SRMR = .04, was a better fit ($\chi^2_{\text{difference}}$ (1) = 166.94, p < .001) than a 1-factor solution, χ^2 (27) = 262.37, p < .001; CFI = .84; TLI = .78; RMSEA = .11; SRMR = .07. This provided further evidence that items in Factor 2, i.e., 4 mentalising items, measured the same latent construct (Table 1).

Study 2: Measure Development, Factor Structure, and Reliability

We made two minor changes to the selected items before examining the internal consistency, factorial validity, and test-retest reliability of the new instrument (Studies 2a & b). First, the item "I sometimes find it difficult to see things from the "other guy's" point of view" was changed to "*I sometimes find it difficult to see things from other people's point of view*" for gender-neutral language. This was reverse worded/scored. Second, the item "I can usually appreciate the other person's viewpoint, even if I do not agree with it" was changed to "*I can usually understand another person's viewpoint, even if it differs from my own*" to further reduce its emotional content without altering its intended meaning. The remaining items were unchanged, as was the 4-point response scale from Study 1. Together, these items formed the Four-Item Mentalising Index (FIMI; see Supplementary Material).

Study 2a. The FIMI was completed online by a large community sample of 1,999 English-speaking adults in the US and UK recruited through Mturk, followed by questions about their age and sex (1286 females, aged 18-80, M_{age} = 35.77, SD_{age} = 12.82 years). There were no other inclusion or exclusion criteria, and there were no attention checks as the study was short in duration. This study was designed to enable i) Multi-group Confirmatory Factor Analyses (MCFA) by sex and ii) a well-powered examination of the link between age, sex, and FIMI scores.

The FIMI had acceptable-to-good internal consistency (ω = .75). To confirm the unidimensional structure of the FIMI, we performed a CFA with all items loading onto 1 factor, which showed that a 1-factor solution provided excellent fit (see Table 2). Further, MCFA tested measurement invariance of the FIMI by sex by comparing 4 models of invariance, each with increasing levels of equality constraint across groups. Configural invariance fit indices were within the critical range and there was a non-significant change between configural, metric, scalar, and strict invariance models in CFI (ΔCFI < .01; Table 3), thereby confirming that the FIMI is invariant to sex.

Further examination of the data showed that each of the FIMI's 4 items had similar levels of variance. Total scores, computed as the sum of the responses to the individual items, ranged between 4 and 16, with higher scores representing greater mentalising ability (see Table 4 for detailed descriptive statistics and Supplementary Table 1 for inter-item correlations). There was no statistically significant association between participant age and total FIMI scores (r = -.01, p = .80), but females (M = 12.60, SD = 2.28) had significantly higher FIMI scores than males (M = 12.08, SD = 2.31), t(1453.58) = 4.44, p < .001, d = 0.21.

Study 2b. Additionally, to determine the test-retest reliability of the FIMI, 116 students (93 females, aged 18-20 years) completed the FIMI in their 1st (M = 12.91, SD = 1.58) and 8th (M = 13.34, SD = 1.61) week at university through an online research participation scheme. FIMI scores were correlated between timepoints (r = .74, p < .001), indicative of acceptable-to-good test-retest reliability. Overall, Study 2 showed that the FIMI has a unidimensional factor structure, measures the same construct in males and females, and has acceptable-to-good reliability.

Study 3: Construct Validity

We examined the validity of the FIMI against two theoretically relevant and psychometrically robust measures: a cognitive mentalising task and a self-report measure of autistic traits (Study 3a). We then validated the FIMI in people who are known to have mentalising difficulties, that is, autistic people (Study 3b).

Study 3a. We tested the relationship between the FIMI and a recently refined version of the Reading the Mind in the Eyes Task (RMET; Olderbak et al., 2015), which requires matching images of the eye region depicting various mental states with one of 4 mental state words. Importantly, this version offers improved validity and reliability to the original task, thereby mitigating potential concerns with its use. We also tested the association between the FIMI and the 28-item Autism-Spectrum Quotient (AQ28; Hoekstra et al., 2011), a widely used measure of autistic traits. We expected that the FIMI would be positively associated with RMET performance, but negatively associated with autistic traits given that mentalising ability is lower in people with many autistic traits (e.g., Livingston, Carr et al., 2019).

Five hundred English-speaking participants (254 females, aged 18-78, M_{age} = 37.29, SD_{age} = 11.74 years), recruited from the US and the UK through Mturk, completed the FIMI (ω = .76), RMET (ω = .79), and AQ28 (ω = .78) in a counterbalanced order, followed by questions about their age and sex. Thirty-four additional participants were excluded because they completed the study twice or failed a simple attention check. As expected, the FIMI was correlated positively with RMET performance ($r = .35$, $p < .001$) and negatively with the AQ28 ($r = -.43$, $p < .001$), particularly its socially relevant subscales (Supplementary Table 2). Females had higher FIMI and RMET scores than male participants, and there was a significant association between age and RMET performance but not age and FIMI scores (see Supplementary Table 3 for descriptives and intercorrelations). Most importantly, we confirmed that the FIMI significantly predicted RMET performance when controlling for age, sex, and AQ28 scores (Table 5). Furthermore, a CFA re-confirmed that the FIMI has a unidimensional factor structure (Table 2).

Study 3b. To further validate the FIMI, we tested whether autistic adults scored lower on the FIMI compared to a non-autistic control group, as should be expected based on previous research (Livingston, Carr et al., 2019). Autistic participants were recruited through the Wales Autism Research Centre, the National Autistic Society, and Autistica's Discover Network, and non-autistic participants were recruited through Prolific. Participants were 102 adults (62 females, aged 18-68, $M_{age} = 41.74$, $SD_{age} = 12.15$ years) with a confirmed clinical diagnosis of autism from a UK healthcare professional (e.g., Psychiatrist) and 183 non-autistic individuals (101 females, aged 18-73, $M_{age} = 40.04$, $SD_{age} = 11.76$ years) who were also from the UK. Groups were matched by age, sex, and general mental ability. As measured by the 10-item Autism-Spectrum Quotient (AQ10; Allison et al., 2012), the autistic group ($M = 8.02$, $SD = 1.96$) had more autistic traits than the non-autistic group ($M = 2.64$, $SD = 1.42$), $t(161.59) = 24.43$, $p < .001$, $d = 3.15$ (see Supplementary Table 4).

All participants completed the FIMI and, as expected, scores were significantly lower in autistic ($M = 9.56$, $SD = 2.69$; ω = 0.73) than non-autistic ($M = 12.57$, $SD = 2.09$; ω = 0.79) participants, $t(169.88) = 9.78$, $p < .001$, $d = 1.25$. Similarly, there was a negative association between the FIMI and AQ10 across groups, $r = -.57$, $p < .001$, and within the autistic group, $r = -.38$ $p < .001$. Replicating the CFA from Study 2a and 3a, the FIMI had a unidimensional factor structure (Table 2). To assess the appropriateness of our groupwise comparisons and the FIMI's use in future autism research, we tested if the FIMI measures the same construct in autistic and non-autistic people using MCFA. This confirmed that the FIMI was invariant to autism at the configural, metric, and scalar level, as evidenced by a non-significant change in CFI (ΔCFI < .01; Table 3). Overall, Study 3 supported the FIMI's validity, given its associations with i) an existing measure of mentalising, ii) autistic traits in the general population, and iii) clinically diagnosed autism.

Discussion

In view of ongoing concerns regarding the conceptual clarity and psychometric validity of existing measures of social cognition (Murphy & Lilienfeld, 2019; Olderbak et al., 2019; Quesque & Rossetti, 2020), we developed the FIMI to selectively measure mentalising ability without it being conflated with other social-cognitive constructs (e.g., empathy), which was possible given our approach to carefully selecting mentalising items for the FIMI (Study 1). This was supported by our analyses showing that the FIMI has a robust unidimensional factor structure and acceptable-to-good internal consistency (Study 2a) and test-retest reliability (Study 2b).

In developing the FIMI, we have neither shown nor claim that mentalising is entirely separable from other social and non-social cognitive constructs. Indeed, mentalising, empathy, and domain-general cognitive processes, like executive function, are likely to be in(ter)dependant (Happé et al., 2017). The FIMI should help to further our understanding of these relationships by facilitating future research on quantifying the extent of the overlap (or lack thereof) between these constructs, and therefore it has potential to contribute to several domains of psychological science. Similarly, because we showed that the FIMI measures the same construct in males and females, its future use will also help to address outstanding questions on sex differences in social cognition and behaviour, particularly in clinical groups with atypical social cognition and uneven sex ratios (e.g., autism, Loomes et al., 2017; social anxiety disorder, Asher et al., 2017). More generally, there has been surprisingly little analysis of sex-based measurement invariance of mentalising and other social-cognition measures, which, as reported in the present study, will be useful to conduct in future research.

A potential limitation of short measures, like the FIMI, is their proneness to measurement error if they are not carefully designed (see Ziegler et al., 2014). Encouragingly, however, the FIMI's internal consistency was acceptable-to-good, in both autistic and non-autistic people, and we found a satisfactory range of scores for a short measure, even within the autistic group (Table 4). Further, Study 3 helped to assuage concerns about the FIMI's reliability and validity. Because the FIMI was significantly associated with performance on an existing measure of mentalising ability, even when accounting for age, sex, and autistic traits, we report strong evidence that it is capturing meaningful individual differences in mentalising. It is particularly noteworthy that the FIMI was more strongly associated with the refined RMET than a well-established measure of autistic traits that has long been associated with performance on the RMET and other mentalising tasks (i.e., Table 5; see Olderbak et al., 2015). Together, these results indicate that, despite its brevity, the FIMI is well designed to quantify individual differences in mentalising.

Our findings also indicate that adults have insight into their social-cognitive processing, thereby addressing recent concerns that social-cognitive abilities may not be measured through self-report scales (e.g., Murphy & Lilienfeld, 2019). The development of the FIMI and Study 3's results showing a link between self-report and objective measures of mentalising, will contribute to and prompt further investigation into this important, under-researched topic. This will not be straightforward as there are ongoing debates about the validity of many cognitive and experimental measures of social cognition (e.g., Livingston, Carr et al., 2019; Quesque & Rossetti, 2020). It will be difficult to investigate people's insight into their social cognition while it remains unclear exactly which processes are measured in experimental cognitive tasks (e.g., Santiesteban et al., 2015). By using the FIMI, it will be possible to reliably measure 'subjective' mentalising ability, which may feed into refining 'objective' cognitive measures of mentalising. This process will particularly benefit from psychometric analyses, as conducted in the present study, which are often missing in social cognition research (see Olderbak et al., 2019). The FIMI therefore could complement rather than replace cognitive measures of social cognition, which, in turn, will serve to improve both self-report and cognitive measures of mentalising.

The use of the FIMI in clinically relevant research and practice was supported by Study 3, which also demonstrated support for the construct validity of the FIMI. That is, we found clear associations between the FIMI and autism. Indeed, the effect size of the difference between autistic and non-autistic people was larger than recent studies using cognitive measures of mentalising in autism (e.g., Murray et al., 2017). These results are consistent with longstanding theory and empirical reports of autism-related mentalising difficulties (e.g., Livingston, Carr et al., 2019). Importantly, we also showed that the FIMI is invariant to autism, thereby supporting its use to measure and/or compare mentalising in autistic and non-autistic people. An association between autistic traits and the FIMI *within* the autistic group also suggests that the FIMI potentially measures clinically meaningful mentalising differences. Although finding autism-related mentalising difficulties may not seem particularly novel, Study 3b is the first to quantify *self-reported* mentalising difficulties in autism *and* the first using an instrument that has formally been shown to be invariant to autism. Together, Study 3 provides validation for the FIMI and fresh support for theories about atypical mentalising in autism.

Moving forward, the FIMI may be useful to gain insight into autistic people's internal experiences and thereby inform diagnosis and social support interventions. In particular, the FIMI may be critical for understanding social difficulties in autistic people who use compensatory strategies, which many existing social-cognitive tasks seem unable to capture (see Livingston, Shah et al., 2019). Given its brevity, the FIMI may also have utility in time-restricted clinical settings. Although the FIMI is shown to be suitable for use in autistic people, further investigation will be required in other clinical groups where atypical mentalising has been reported (e.g., social anxiety disorder). More broadly, the FIMI may prove useful in large-scale studies, including online and/or population-based studies, towards addressing important questions about developmental trajectories, genetic underpinnings, and individual differences pertaining to (a)typical mentalising (see Livingston & Happé, in press). A limitation of the current study is that we only explored age- and sex- related mentalising differences while developing the FIMI. However, given the FIMI's brevity, it will be feasible to use it in future large-scale studies to address outstanding questions on other socio-demographic predictors of mentalising (e.g., socioeconomic status; Pluck et al., in press).

Summary & Conclusion

In summary, we report the development and validation of the FIMI, the first self-report measure of mentalising. To develop this measure, we drew from an existing self-report measure of social cognition to derive 4 items putatively measuring mentalising ability (Study 1). Following changes to these items to construct the FIMI, we demonstrated its internal consistency, unidimensional factor structure, invariance to participant sex, and test-retest reliability (Study 2). Finally, we confirmed relationships between the FIMI and existing measures of mentalising and autism, suggesting it has good construct validity (Study 3).

Together, our studies show that the FIMI is a conceptually and psychometrically robust measure of mentalising, suitable for use in autistic and non-autistic adults. The FIMI will help to address limitations in previous mentalising research and is likely to be useful for clinicians interested in (a)typical social cognition.

References

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "red flags" for autism screening: The Short Autism Spectrum Quotient and Short Quantitative Checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(2), 202-212.

Asher, M., Asnaani, A., & Aderka, I. M. (2017). Gender differences in social anxiety disorder: A review. *Clinical Psychology Review*, *56*, 1-12.

Bernaards, C. A., & Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement*, *65*(5), 676-696. https://cran.r-project.org/web/packages/GPArotation/index.html

Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researchers. *Communication Methods and Measures*, *12*(1), 25-44.

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52-64.

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

Happé, F., Cook, J. L., & Bird, G. (2017). The structure of social cognition: In(ter)dependence of sociocognitive processes. *Annual Review of Psychology*, *68*, 243-267.

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

Hoekstra, R. A., Vinkhuyzen, A. A., Wheelwright, S., Bartels, M., Boomsma, D. I., Baron-Cohen, S., Posthuma, D., & Van Der Sluis, S. (2011). The construction and validation of an abridged version of the Autism-Spectrum Quotient (AQ-Short). *Journal of Autism and Developmental Disorders*, *41*(5), 589-596.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2018). semTools: Useful tools for structural equation modeling. https://CRAN. Rcran.r-project.org/package=web/packages/semTools/index.html

Lever, A. G., & Geurts, H. M. (2016). Age-related differences in cognition across the adult lifespan in Autism Spectrum Disorder. *Autism Research*, *9*(6), 666-676.

Livingston, L. A., & Happé, F. (in press). Understanding atypical social behaviour through social cognitive theory: Lessons from autism. In, H. J., Ferguson, V. E. A., Brunsdon & E. E. F. Bradford (Eds.), *The cognitive basis of social interaction across the lifespan*. Oxford University Press.

Livingston, L. A., Carr, B., & Shah, P. (2019). Recent advances and new directions in measuring theory of mind in autistic adults. *Journal of Autism and Developmental Disorders*, *49*(4), 1738-1744.

Livingston, L. A., Shah, P., & Happé, F. (2019). Compensatory strategies below the behavioural surface in autism: a qualitative study. *The Lancet Psychiatry*, *6*(9), 766-777.

Loomes, R., Hull, L., & Mandy, W. (2017). What is the male-to-female ratio in Autism Spectrum Disorder? A systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, *56*(6), 466-474.

McGrath, R. E. (2005). Conceptual complexity and construct validity. *Journal of Personality Assessment*, *85*(2), 112-124.

Morrison, K. E., Pinkham, A. E., Kelsven, S., Ludwig, K., Penn, D. L., & Sasson, N. J. (2019). Psychometric evaluation of social cognitive measures for adults with autism. *Autism Research*, *12*(5), 766-778.

Murphy, B. A., & Lilienfeld, S. O. (2019). Are self-report cognitive empathy ratings valid proxies for cognitive empathy ability? Negligible meta-analytic relations with behavioral task performance. *Psychological Assessment*, *31*(8), 1062-10691072.

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 Research*, *10*(6), 1120-1132.

Navarro, E., Goring, S. A., & Conway, A. R. A. (in press). The relationship between theory of mind and intelligence: A formative g approach. *Intelligence.*

Olderbak, S., Geiger, M., & Wilhelm, O. (2019). A call for revamping socio-emotional ability research in autism. *Behavioral and Brain Sciences*, *42,* e104.

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: toward a brief form for research and applied settings. *Frontiers in Psychology*, *6*, 1503.

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

Preckel, K., Kanske, P., & Singer, T. (2018). On the interaction of social affect and cognition: empathy, compassion and theory of mind. *Current Opinion in Behavioral Sciences*, *19*, 1-6.

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

Reniers, R. L., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment*, *93*(1), 84-95.

Revelle, W. (2018). psych: Procedures for personality and psychological research., psychometric, and personality research. Northwestern University, Evanston. https://cran.rprojectr-project.org/web/packages/psych/index.html

Rosseel. Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://cran.r-project.org/web/packages/lavaan/index.html

Rueda, P., Fernández-Berrocal, P., & Baron-Cohen, S. (2015). Dissociation between cognitive and affective empathy in youth with Asperger Syndrome. *European Journal of Developmental Psychology*, *12*(1), 85-98.

Santiesteban, I., Shah, P., White, S., Bird, G., & Heyes, C. (2015). Mentalizing or submentalizing in a communication task? Evidence from autism and a camera control. *Psychonomic Bulletin & Review*, *22*, 844-849.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis*. *A Festschrift for Heinz Neudecker* (pp.233-247). Kluwer Academic Publishers.

Shah, P., Livingston, L. A., Callan, M. J., & Player, L. (2019). Trait autism is a better predictor of empathy than alexithymia. *Journal of Autism and Developmental Disorders*, *49*(10), 3956-3964.

Svetina, D., Rutkowski, L., & Rutkowski, D. (2019). Multiple-group invariance with categorical outcomes using updated guidelines: an illustration using M plus and the Lavaan/semTools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 111-130.

Taylor, E. C., Livingston, L. A., Callan, M. J., & Shah, P. (2019). Divergent contributions of autistic traits to social psychological knowledge. *Proceedings of the National Academy of Sciences*, *116*(51), 25378-25379.

Warnell, K. R., & Redcay, E. (2019). Minimal coherence among varied theory of mind measures in childhood and adulthood. *Cognition*, *191*, 103997.

White, S. J., Coniston, D., Rogers, R., & Frith, U. (2011). Developing the Frith-Happé Animations: A quick and objective test of Theory of Mind for adults with autism. *Autism Research*, *4*(2), 149-154.

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81*(4), 1014-1045.

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales–Five misunderstandings and ways to overcome them. *Journal of Individual Differences*, *35*(4), 185-189.

Tables

Table 1. *Factor Loadings for Non-emotional Items of the QCAE (Studies 1a and 1b).*

Item	EFA (Study 1a)		CFA (Study 1b)	
	Factor 1	Factor 2	Factor 1	Factor 2
I can easily tell if someone else wants to enter a conversation (15)	.58	.10	.67	--
I can pick up quickly if someone says one thing but means another (16)	.72	-.10	.62	--
I can sense if I am intruding, even if the other person does not tell me (24)	.57	.09	.56	--
I can easily work out what another person might want to talk about (25)	.67	.05	.65	--
I am good at predicting what someone will do (27)	.56	.00	.69	--
I find it easy to put myself in somebody else's shoes (18)	.04	.73	**--**	.69
I sometimes find it difficult to see things from the "other guy's" point of view (1, Reversed)	-.12	.65	**--**	.48
I sometimes try to understand my friends better by imagining how things look from their perspective (4)	.13	.57	**--**	.59
I can usually appreciate the other person's viewpoint, even if I do not agree with it (28)	.06	.51	**--**	.66

Note. Potential mentalising items are presented in bold font (note the wording is amended for the FIMI, see Table 2). QCAE: Questionnaire of Cognitive and Affective Empathy; EFA: Exploratory Factor Analysis; CFA: Confirmatory Factor Analysis. Original QCAE item numbers are in parentheses.

Table 2. *Factor Loadings for the FIMI Items Across Studies using Confirmatory Factor Analysis (CFAs) of a Single-Factor Model.*

Item	Study 2a	Study 3a	Study 3b
1. I find it easy to put myself in somebody else's shoes	.80	.79	.79
2. I sometimes find it difficult to see things from other people's point of view (Reversed)	.60	.45	.77
3. I sometimes try to understand my friends better by imagining how things look from their perspective	.60	.72	.62
4. I can usually understand another person's viewpoint, even if it differs from my own	.60	.69	.73

Note. All CFAs indicated that the FIMI had a unidimensional factor structure, with excellent fit indices across all the studies (Study 2a: χ^2 (2) = 8.41, *p* = .015; CFI > .99; TLI = .99; RMSEA = .04; SRMR= .01; Study 3a: χ^2 (2) = 9.50, *p* = .009; CFI = .99, TLI = .96, RMSEA = .09, SRMR = .03; Study 3b: χ^2 (2) = 2.72, *p* = .26; CFI > .99, TLI = .99, RMSEA = .04, SRMR = .02).

Table 3. *Measurement Invariance of the FIMI by Sex (Study 2a) and Autism (Study 3b).*

		χ^2	df	$\Delta\chi^2$	CFI	ΔCFI	TLI	RMSEA	SRMR
Sex	Configural	9.29	4		>.99		.99	.04	.01
	Metric	10.95	7	1.66	>.99	<.01	>.99	.02	.01
	Scalar	12.74	10	1.79	>.99	<.01	>.99	.02	.02
	Strict	19.51	14	6.77	>.99	<-.01	>.99	.02	.02
Autism	Configural	13.84	4		.97		.90	.13	.03
	Metric	18.57	7	4.73	.96	<-.01	.93	.11	.05
	Scalar	24.10	10	5.53	.95	<-.01	.94	.10	.05
	Strict	94.39	14	70.30	.71	<-.24	.76	.20	.12

Note: χ^2: chi-square goodness of fit test; df: degrees of freedom; CFI: Comparative Fit Index; TLI: Tucker-Lewis Index; RMSEA: Root Mean Squared Error of Approximation; SRMR: Standardised Root Mean Squared Residual. Strict invariance, which tests equivalence of residuals between groups, was neither expected nor obtained given the generally higher level of variance typically found in autistic compared to non-autistic groups.

Table 4. *Detailed Descriptive Statistics of Total FIMI Score by Study*

	Study						
Statistic	2a	2b Week 1	2b Week 8	3a	3b Overall	3b Autistic	3b Non-autistic
Mean	12.39 (.05)	12.91 (.15)	13.34 (.15)	12.25 (.11)	11.49 (.16)	9.56 (.27)	12.57 (.15)
Variance	5.29	2.50	2.59	5.79	7.46	7.22	4.37
SD	2.30	1.58	1.61	2.41	2.73	2.69	2.09
Range	12	8	7	12	12	12	12
IQR	3	2	3	3	3	3	3
Skewness	-0.36 (.06)	-0.17 (.23)	-0.03 (.23)	-0.22 (.11)	-0.52 (.14)	-0.09 (.24)	-0.37 (.18)
Kurtosis	0.12 (.11)	0.09 (.45)	-0.58 (.45)	0.11 (.22)	0.09 (.29)	-0.41 (.47)	0.75 (.36)
Median	12	13	13	12	12	10	12
MAD	2.97	1.48	1.48	1.48	2.97	2.97	1.48
Robust							

Note. SD: Standard Deviation; IQR: Interquartile Range; MAD: Median Absolute Deviation. Standard Errors are in parentheses.

Table 5. *Multiple Regression Analysis of the FIMI, AQ28, Age, and Sex Predicting RMET Performance (Study 3a).*

	RMET							
Predictors	*B*	*SE B*	*β*	*t*	*p*	sr^2	95% BCa CI Lower	95% BCa CI Upper
FIMI	.34	.05	.30	6.55	<.001	.07	0.24	0.44
AQ28 (Autistic Traits)	-.02	.01	-.05	-1.15	.25	.00	-0.04	0.01
Age	.04	.01	.18	4.46	<.001	.03	0.03	0.06
Sex	-.49	.23	-.09	-2.16	.031	.01	-0.94	-0.05

Note. Overall Model: $F(4, 495) = 24.71$, $p < .001$, $R^2 = .17$. Sex is coded as Males = 1, Females = 0. 95% Bias corrected and accelerated bootstrapped confidence intervals (95% BCa CI) are reported. RMET: Reading the Mind in the Eyes Test; FIMI: Four-Item Mentalising Index; AQ28: 28-item Autism-Spectrum Quotient.

Supplementary Materials

Clutterbuck, R. A., Callan, M. J., Taylor, E. C., Livingston, L. A., & Shah, P. Development and validation of the Four-Item Mentalising Index (FIMI). *Psychological Assessment*.

Confirmatory Factor Analysis (CFA) using Weighted Least Square Mean and Variance adjusted (WLSMV) estimation

Repeating CFAs in Study 1a data using WLSMV revealed an almost identical pattern of results as those using Maximum Likelihood (ML) estimation. Separating 9 non-emotional from 22 emotional ($\chi^2_{difference}$ (1) = 57.74, $p < .001$), and 4 mentalising from 27 remaining QCAE items ($\chi^2_{difference}$ (1) = 126.24, $p < .001$) in two-factor models had significantly better fit than a unifactorial model of all QCAE 31 items. All other repeated CFAs with WLSMV estimation produced similar results to ML, in terms of factor loadings, model fit indices, and measurement invariance across Studies 1-3 (see Supplementary Tables 5-7).

Supplementary Table 1

Inter-item Correlations of the Four-Item Mentalising Index (Study 2a).

	M(SD)	Skew	Kurtosis	Q1	Q2	Q3	Q4
Q1. I find it easy to put myself in somebody else's shoes	3.13 (0.78)	-.71	.22	-			
Q2. I sometimes find it difficult to see things from other people's point of view (Reversed)	2.75 (0.89)	-.15	-.80	.49*	-		
Q3. I sometimes try to understand my friends better by imagining how things look from their perspective	3.30 (0.70)	-.87	.83	.48*	.33*	-	
Q4. I can usually understand another person's viewpoint, even if it differs from my own	3.21 (0.69)	-.62	.43	.47*	.36*	.38*	-

Note. *p < .001

Supplementary Table 2

Correlations Between the FIMI and AQ28 Subscales (Study 3a).

Measure	*M(SD)*	1	2	3	4	5	6
1. FIMI	12.25(2.41)	-					
2. AQ28 Total	65.27(9.10)	-.43**	-				
AQ28 Subscales							
3. Social Skills	19.20(4.84)	-.22**	.79**	-			
4. Routine	7.48(1.82)	-.16**	.48**	.41**	-		
5. Switching	8.91(2.06)	-.22**	.61**	.46**	.32**	-	
6. Imagination	16.55(3.77)	-.59**	.66**	.31**	.17**	.31**	-
7. Numbers & Patterns	13.14(3.36)	.05	.19**	-.15*	-.20**	-.15**	-.06

Note. Pearson's *r* values presented. AQ28: 28-item Autism-Spectrum Quotient. FIMI: Four-Item Mentalising Index. *$p < .05$, **$p < .001$. We thank an anonymous reviewer for suggesting this analysis, which follows recent research on divergent contributions of autistic traits to social cognition (Taylor et al., 2019).

Supplementary Table 3

Descriptive Statistics and Intercorrelations (Study 3a).

Measure	*M(SD)*	1	2	3	4	5
1. AQ28	65.27 (9.10)	-				
2. RMET	6.33 (2.76)	-.20**	-			
3. FIMI	12.25 (2.41)	-.43**	.35**	-		
4. Age	37.29 (11.74)	-.07	.22**	.08	-	
5. Sex	0.49 (0.50)	.05	-.14*	-.10*	-.09*	-

Note. Pearson's *r* values presented. AQ28: 28-item Autism-Spectrum Quotient. RMET: Reading the Mind in the Eyes Test. FIMI: Four-Item Mentalising Index. Sex is coded as Males = 1, Females = 0. **p* < .05, ***p* < .001.

Supplementary Table 4

Participant Characteristics and Matching of the Autistic and Non-Autistic Groups (Study 3b).

	Autistic (n = 102)	Non-Autistic (n = 183)	Range	Group Difference
	M(SD)	*M(SD)*		
Age	41.74(12.15)	40.04(11.76)	18–73	t (203.29) = 1.14, p = .25, d = 0.14
General Mental Ability	9.02(3.77)	8.58(3.07)	4–16	t (175.99) = 1.01, p = .31, d = 0.13
n Male, Female	40, 62	82, 101		$\chi 2(1)$ = 0.84, p = .36, Φ = 0.05

Note. General Mental Ability was quantified using the International Cognitive Ability Resource, which has previously demonstrated concurrent validity with in-person assessments (see Condon & Revelle, 2014). Cohen's *d* was interpreted as 0.2 = small, 0.5 = medium, and 0.8 = large effects. Phi *Φ* is interpreted as 0.1 = small, 0.3 = medium, and 0.5 = large effects.

Supplementary Table 5

Factor Loadings for Non-emotional Items of the QCAE (Study 1b) using Confirmatory Factor Analysis (CFA) with WLSMV Estimation of a Two-Factor Model.

Item	Factor 1	Factor 2
I can easily tell if someone else wants to enter a conversation (15)	.72	--
I can pick up quickly if someone says one thing but means another (16)	.68	--
I can sense if I am intruding, even if the other person does not tell me (24)	.65	--
I can easily work out what another person might want to talk about (25)	.72	--
I am good at predicting what someone will do (27)	.77	--
I find it easy to put myself in somebody else's shoes (18)	**--**	.76
I sometimes find it difficult to see things from the "other guy's" point of view (1, Reversed)	**--**	.49
I sometimes try to understand my friends better by imagining how things look from their perspective (4)	**--**	.67
I can usually appreciate the other person's viewpoint, even if I do not agree with it (28)	**--**	.77

Note. Potential mentalising items are presented in bold font (note the wording is amended for the FIMI, see Supplementary Table 5). QCAE: Questionnaire of Cognitive and Affective Empathy. Original QCAE item numbers are shown in parentheses. CFA with WLSMV estimation showed that a 2-factor solution, χ^2 (26) = 126.02, p < .001; CFI = .96; TLI = .95; RMSEA = .08; SRMR = .05, was a better fit ($\chi^2_{\text{difference}}$(1) = 120.35, p < .001) than a 1-factor solution, χ^2 (27) = 422.26, p < .001; CFI = .86; TLI = .81; RMSEA = .15; SRMR = .09.

Supplementary Table 6

Factor Loadings for the FIMI Items Across Studies using Confirmatory Factor Analysis (CFAs) with WLSMV Estimation of a Single-Factor Model.

Item	Study 2a	Study 3a	Study 3b
5. I find it easy to put myself in somebody else's shoes	.87	.87	.84
6. I sometimes find it difficult to see things from other people's point of view (Reversed)	.65	.52	.83
7. I sometimes try to understand my friends better by imagining how things look from their perspective	.67	.79	.68
8. I can usually understand another person's viewpoint, even if it differs from my own	.67	.75	.78

Note. Study 2a: χ^2 (2) = 11.13, *p* = .004; CFI > .99, TLI = .99, RMSEA = .05, SRMR = .01; Study 3a: χ^2 (2) = 11.40, *p* = .003; CFI = .99, TLI = .98, RMSEA = .10 SRMR = .03; Study 3b: χ^2 (2) = 3.58, *p* = .17; CFI > .99, TLI > .99, RMSEA = .05, SRMR = .02.

Supplementary Table 7

Measurement Invariance of the FIMI by Sex (Study 2a) and Autism (Study 3b) using WLSMV Estimation Following Svetina et al. (2019; see also Wu & Estabrook, 2016).

		Scaled χ^2	df	$\Delta\chi^2$	Scaled CFI	Δ Scaled CFI	Scaled TLI	Scaled RMSEA	SRMR
Sex	Configural	12.71	4		>.99		.99	.05	.01
	Equal Thresholds	26.74	8	10.42	>.99	<-.01	.99	.05	.01
	Equal Thresholds and Loadings	22.55	11	1.39	>.99	<.01	>.99	.03	.01
Autism	Configural	16.25	4		.98		.95	.15	.05
	Equal Thresholds	21.68	8	4.50	.98	<-.01	.97	.11	.05
	Equal Thresholds and Loadings	22.26	11	2.69	.99	<.01	.98	.09	.05

Note. χ^2: chi-square goodness of fit test; df: degrees of freedom; CFI: Comparative Fit Index; TLI: Tucker-Lewis Index; RMSEA: Root Mean Squared Error of Approximation; SRMR: Standardised Root Mean Squared Residual. The $\Delta\chi^2$ represent scaled chi-squared differences following Satorra (2000).

Supplementary References

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52-64.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis*. *A Festschrift for Heinz Neudecker* (pp.233-247). Kluwer Academic Publishers.

Svetina, D., Rutkowski, L., & Rutkowski, D. (2019). Multiple-group invariance with categorical outcomes using updated guidelines: an illustration using M plus and the Lavaan/semTools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 111-130.

Taylor, E. C., Livingston, L. A., Callan, M. J., & Shah, P. (2019). Divergent contributions of autistic traits to social psychological knowledge. *Proceedings of the National Academy of Sciences*, *116*(51), 25378-25379.

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81*(4), 1014-1045.