Machine learning and deep learning systems for automated measurement of ‘advanced’ theory of mind

Devine, R.T.; Kovatchev, Venelin; Grumley Traynor, Imogen; Smith, Phillip; Lee, Mark

DOI:
10.1037/pas0001186

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Assessment

ISSN: 1040-3590

2023, Vol. 35, No. 2, 165–177
https://doi.org/10.1037/pas0001186

Machine Learning and Deep Learning Systems for Automated Measurement of "Advanced" Theory of Mind: Reliability and Validity in Children and Adolescents

Rory T. Devine[1], Venelin Kovatchev[1], Imogen Grumley Traynor[1], Phillip Smith[2], and Mark Lee[2]
[1] School of Psychology, University of Birmingham
[2] School of Computer Science, University of Birmingham

Understanding individual differences in theory of mind (ToM; the ability to attribute mental states to others) in middle childhood and adolescence hinges on the availability of robust and scalable measures. Open-ended response tasks yield valid indicators of ToM but are labor intensive and difficult to compare across studies. We examined the reliability and validity of new machine learning and deep learning neural network automated scoring systems for measuring ToM in children and adolescents. Two large samples of British children and adolescents aged between 7 and 13 years (Sample 1: $N = 1{,}135$, $M_{age} = 10.22$ years, $SD = 1.45$; Sample 2: $N = 1{,}020$, $M_{age} = 10.36$ years, $SD = 1.27$) completed the silent film and strange stories tasks. Teachers rated Sample 2 children's social competence with peers. A single latent-factor explained variation in performance on both the silent film and strange stories task (in Sample 1 and 2) and test performance was sensitive to age-related differences and individual differences within each age-group. A deep learning neural network automated scoring system trained on Sample 1 exhibited interrater reliability and measurement invariance with manual ratings in Sample 2. Validity of ratings from the automated scoring system was supported by unique positive associations between ToM and teacher-rated social competence. The results demonstrate that reliable and valid measures of ToM can be obtained using the new freely available deep learning neural network automated scoring system to rate open-ended text responses.

Public Significance Statement
Children differ from one another in their understanding of other people's thoughts and feelings (called "theory of mind") and these differences matter for children's social lives. We open up new opportunities for research by showing that machine learning and deep learning algorithms can automatically score children's responses to open-ended tests of theory of mind, making large-scale, robust studies possible without the need for labor-intensive scoring by trained researchers.

Keywords: theory of mind, middle childhood, adolescence, machine learning, deep learning

Supplemental materials: https://doi.org/10.1037/pas0001186.supp

Rory T. Devine https://orcid.org/0000-0002-3710-7878

This project was funded through a Wellcome Trust Grant 215006/Z/18/Z to Rory T. Devine. The authors wish to thank Irene Luque Aguilera and Tom Willetts for their assistance with data collection and scoring. We also want to thank the children and teachers who participated in our research.

Rory T. Devine and Venelin Kovatchev made an equal contribution to the article.

Rory T. Devine played lead role in conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, writing of original draft and writing of review and editing. Venelin Kovatchev played lead role in software and data curation and supporting role in conceptualization, formal analysis, methodology and writing of original draft. Imogen Grumley Traynor played supporting role in investigation, methodology and project administration. Phillip Smith played supporting role in conceptualization, formal analysis, funding acquisition, methodology and supervision. Mark Lee played supporting role in funding acquisition, methodology, project administration, software, supervision and writing of review and editing and equal role in conceptualization.

This study was preregistered on the open science framework (OSF): https://osf.io/rxyfh.

Test materials and scoring materials are available on the OSF: https://osf.io/8x73r/.

Correspondence concerning this article should be addressed to Rory T. Devine, School of Psychology, University of Birmingham, 52 Prichatts Road, Edgbaston, Birmingham B15 2SB, United Kingdom. Email: R.T.Devine@bham.ac.uk

The ability to tune into others' mental states, called "theory of mind" (ToM) or "mindreading", has garnered substantial interest in psychology since the early 1980s. Curiosity about ToM can perhaps be explained by three trends. First, there is evidence of ongoing development in ToM beyond early childhood across middle childhood and adolescence (Weimer et al., 2021) and of early emerging and stable individual differences in ToM (Devine, 2021). Second, rather than being limited to conditions such as autism, ToM difficulties cross-cut a range of mental health and neurodevelopmental conditions (Cotter et al., 2018). Third, individual differences in ToM performance are meaningful: Children who excel at tests of ToM are more popular and prosocial than their peers (Imuta et al., 2016; Slaughter et al., 2015). Continued progress in understanding ToM hinges on the availability of robust measures. Despite more than 40 years of research, relatively few studies have examined the psychometric properties of ToM tests (Osterhaus & Bosacki, 2022). The aim of the present study was to examine the reliability and validity of machine learning and deep learning automated scoring systems for scalable, reproducible measurement of ToM in children and adolescents.

Measuring "Advanced" ToM

In early childhood ToM is typically measured using brief forced-choice response tasks. The false belief task is sensitive to developmental differences in preschool children (Wellman et al., 2001), can be combined with other measures of false belief understanding to capture individual differences (Hughes & Devine, 2015), and demonstrates good test-retest reliability (Hughes et al., 2000). However, standard versions of the false belief task exhibit ceiling effects beyond early childhood (Peterson et al., 2012), making it unsuitable for use in middle childhood and adolescence (Devine, 2021).

Extending ToM research beyond early childhood has led to the creation of a variety of tests (e.g., Osterhaus & Bosacki, 2022). These "advanced" ToM tasks vary in three ways. First, advanced ToM tasks incorporate diverse stimuli including short vignettes (Lagattuta & Kramer, 2021a, 2021b), film clips (Devine & Hughes, 2016), animations (Livingston et al., 2021), and static images (Meinhardt-Injac et al., 2020). Second, advanced ToM tasks target different skills including emotion recognition (Cassels & Birch, 2014), perspective taking (Tamnes et al., 2018), recursive reasoning (Osterhaus & Koerber, 2021), and predicting, interpreting or explaining behavior (White et al., 2009). Third, advanced ToM tasks use forced-choice (Osterhaus & Koerber, 2021) and open-ended response formats (Devine & Hughes, 2013). Despite task differences, studies show age-related gains in performance across a range of advanced ToM tasks in middle childhood and adolescence (e.g., Meinhardt-Injac et al., 2020; Osterhaus & Koerber, 2021). However, research on the psychometric properties of "advanced" ToM tests is rare (e.g., Osterhaus & Bosacki, 2022).

Studies using the Strange Stories (Happé, 1994) and Silent Film tasks (Devine & Hughes, 2013) show that these tasks are sensitive to developmental and individual differences in ToM in middle childhood and adolescence (Devine & Hughes, 2016). In the Strange Stories task (White et al., 2009), participants are presented with short text vignettes and answer an open-ended question after each story. In the Silent Film task, participants watch and answer questions about scenarios depicted in short silent film clips. The vignettes and clips each involve perspective differences between the participant and characters and depict instances of misunderstanding, deception, misdirection, pretense, and surprise across a range of different contexts (Table S1). Participants are required to explain a target's actions by inferring their intentions, beliefs, and emotions to varying degrees (Table S1). Despite the heterogeneity of items, responses are rated using a coding scheme where success reflects the ability to explain a character's behavior in a given context by tuning in to the character's mental states (Devine & Hughes, 2016; White et al., 2009). These tasks are sensitive to age-related differences in 7- to 13-year-old children's performance and capture individual differences within each age-group (Devine & Apperly, 2022; Lecce et al., 2017).

Scores on the Silent Film and Strange Stories tasks appear to be reliable and valid indicators of ToM in middle childhood and early adolescence. Confirmatory factor analyses show that items from each task load onto a single latent factor despite differences in stimuli (i.e., text vignettes *vs.* film clips) and item content (Devine & Hughes, 2013; Devine & Hughes, 2016). The combined tasks are more reliable than either task alone in terms of internal consistency (>.70; Devine & Hughes, 2016) and are more sensitive to individual and age-related differences in performance than either task alone (Devine & Hughes, 2013). Latent factor scores exhibit good 1 month test–retest reliability ($r > .80$) across a wide range of ability levels (Devine & Hughes, 2016).

At least two lines of evidence support the validity of the Silent Film and Strange Stories tasks as measures of ToM. First, convergent validity is supported by associations between performance on these two tasks and scores on the Triangles Task (Castelli et al., 2000), another widely used measure of ToM (Devine et al., 2016). Longitudinal data indicate that earlier performance on a battery of false belief tasks (the most widely used measures of ToM) at age 6 predicts later scores on both the silent film and strange stories task at age 10 suggesting these tasks measure the same construct (Devine et al., 2016). These concurrent and longitudinal associations hold even when individual differences in verbal ability and executive function are considered, suggesting that test performance on the Silent Film and Strange Stories tasks does not simply reflect differences in these abilities (Devine et al., 2016).

Second, scores on the silent film and strange stories tasks exhibit criterion validity. According to the "social individual differences" account of ToM, individual differences in tasks that measure ToM should be associated with children's social competence (i.e., their ability to build, manage, and maintain social relationships; Apperly, 2012). Note that ToM is not viewed as synonymous with social competence as many aspects of social competence do not rely on making inferences about others' mental states (e.g., Lecce & Devine, 2021). However, links with real-world social outcomes provide support for the claim that individual differences in ToM test performance are meaningful (Lecce & Devine, 2021). The absence of associations between a purported measure of ToM and indices of social competence would undermine the validity of that task. Meta-analyses indicate positive associations between children's ToM and prosocial behavior ($r = .19$) and between ToM and peer acceptance ($r = .19$; Imuta et al., 2016; Slaughter et al., 2015). On this basis, it is expected that measures of individual differences in ToM should exhibit similar strength associations with measures of social competence. Given that both social competence and ToM have been linked with individual differences in verbal ability, socioeconomic status, gender, and a range of developmental conditions (e.g., Bratsch-Hines et al., 2020; Weimer et al., 2021), it is also important to consider the impact of these confounds when attempting to

establish unique associations between ToM and children's social competence. Supporting the validity of the Strange Stories and Silent Film tasks, scores on these tests are moderately associated with teachers' ratings of children's social competence at school (Devine & Apperly, 2022), even when confounding variables such as language and socioeconomic status are considered.

One challenge for measurement is that there is no agreed upon account of ToM development beyond early childhood making age differences in test scores difficult to interpret (Hughes & Devine, 2015). It is unclear whether the latent ability measured by the Silent Film and Strange Stories tasks reflects children's verbosity rather than their ability to reason about others' minds. One possibility is that the latent factor captures individual differences in response length. Furthermore, given that both tasks require children to explain others' behavior, it is unclear whether the latent factor captures variation in children's tendency to refer to others' mental states, regardless of whether these are related to the context of the film clip or vignette. Our first aim was therefore to test the factor structure of the Silent Film and Strange Stories tasks in a large sample of children and cross-validate the factor structure in a second sample. We examined if items loaded onto a single factor when answer length was considered and tested an alternative rating scheme capturing children's references to others' mental states, regardless of context.

Open-Ended Response Tests: Challenges and Opportunities

Although the Silent Film and Strange Stories tasks can be administered in group settings and take approximately 15 min to complete (Devine & Hughes, 2016), open-ended response tasks, are viewed as unsuitable for large-scale research (Livingston et al., 2021). Both tasks generate open-ended text responses, which are later manually scored by trained coders (Devine & Hughes, 2016; White et al., 2009). Although high-levels of interrater reliability on these measures are attainable (Devine & Apperly, 2022), training and coding open-ended text is labor intensive and time consuming (Iliev et al., 2015). Differences between coders in item interpretation and deviation from training can hamper reliability (i.e., coders may assign different scores to the same data) and undermine test validity (i.e., scores may reflect different dimensions of ability; Girard & Cohn, 2016; Walker & Göçer Şahin, 2020).

Arguably, open-ended response tasks yield insight into ToM that forced-choice tasks do not. Open-ended tasks are more ecologically valid than forced-choice formats because participants are not presented with cues to mentalize about others (e.g., Cassels & Birch, 2014). Forced-choice versions of social cognition tasks are easier than open-ended versions (Betz et al., 2019) and less closely associated with relevant social outcomes (Cassels & Birch, 2014). An automated scoring system for open-ended responses to the Silent Film and Strange Stories tasks has the potential to facilitate reliable, large-scale studies. Our second aim was therefore to examine the reliability of machine learning and deep learning automated scoring systems for rating open-ended responses to these tasks.

Machine Learning, Deep Learning, and Psychological Assessment

Machine learning algorithms identify patterns in data to make predictions beyond the initial data (Alpaydin, 2016; Chen & Wojcik, 2016) and have been applied to process textual data automatically (Iliev et al., 2015). Using "gold standard" data from human coders (e.g., a rating or score), supervised machine learning algorithms can identify regularities in data (e.g., an open-ended text response) and use these to generalize beyond the provided examples to score new data. Machine learning has been applied in psychology to make judgments about personal attributes based on facial images (Kosinski, 2021) and rate depressive symptoms using speech, video, and textual data from a diagnostic interview (Victor et al., 2019). Machine learning has also been identified as a tool for rating text responses to automate labor-intensive scoring in psychological assessments (Iliev et al., 2015) and has been leveraged to create automatic scoring systems for tests of autobiographical memory (Takano et al., 2018) and divergent thinking (LaVoie et al., 2020).

Traditional machine learning algorithms, such as support vector machines (SVM), make classifications using particular predefined features (Iliev et al., 2015). Feature extraction requires researchers to identify features manually in advance (e.g., word frequency, punctuation markers). In contrast, deep learning neural networks, such as bidirectional long short-term memory (BiLSTM; Graves & Schmidhuber, 2005) and Transformer neural networks such as Bidirectional Encoder Representation from Transformers or "BERT" and "distilled" "DistilBERT" (Sanh et al., 2020), do not require feature engineering. Deep learning neural networks exhibit better performance than traditional machine learning algorithms in a variety of language processing tasks such as paraphrase identification (Kovatchev et al., 2019). We tested whether machine learning and deep learning automated scoring systems can provide an end-to-end solution for scoring ToM whereby the system takes children's open-ended responses to the Silent Film and Strange Stories tasks and returns assigned scores for each item, replacing the need for manual scoring.

We trained and tested the machine learning algorithms and deep learning neural networks for scoring the Silent Film and Strange Stories tasks in 1,135 7- to 13-year-old children (Kovatchev et al., 2020). Deep learning neural networks, such as BiLSTM and DistilBERT, outperformed the SVM algorithm (trained on basic linguistic features) on standard evaluation metrics (i.e., system accuracy and Macro-F1). These scoring systems are promising but it is not yet clear whether high performance on evaluation metrics for classification algorithms correspond with good quality data for psychological research. If the same data were scored twice (i.e., by manual and automatic scoring), then evidence of measurement noninvariance (i.e., differences in factor structure, loadings or thresholds) would signal that the automatic scoring system was applying the scoring system differently to the same data (Walker & Göçer Şahin, 2020). If latent factor scores derived from manual and automatic scoring were not related to the same outcomes, then this would indicate that the automatic scoring system measured a different variable. Our third aim was to investigate whether automated scoring systems capture the same latent factor as manual ratings and whether automated scores yield valid measures of children's ToM in an independent sample.

Summary of Aims and Hypotheses

The overarching aim of the present study was to examine the reliability and validity of machine learning and deep learning automated scoring systems for measuring ToM in children and

adolescents. Our first aim was to test and cross-validate the latent factor structure of the Silent Film and Strange Stories tasks. We predicted that a single latent factor would provide a good fit to the data and that items would load onto this latent factor even when answer length was considered. Our second aim was to examine the reliability of new machine learning and deep learning automated scoring systems for the Silent Film and Strange Stories tasks by comparing item-level reliability of automated scores with manual ratings and testing for measurement invariance (Brown, 2015). If automated scoring captures the same latent ability as manual ratings, then manual and automated scores will exhibit similar factor structure and loadings (i.e., the latent factors are defined similarly) and equal item thresholds (i.e., similar levels of underlying ToM ability are associated with obtaining ratings of 0, 1, or 2; Walker & Göçer Şahin, 2020). Our third aim was to test the validity of scores derived from the best performing automated scoring system by comparing the associations between manual and automated ToM ratings and teacher-rated social competence. If automated scoring is valid, then automated and manual ToM scores will exhibit equivalent associations with social competence.

Method

Participants

Sample 1 comprised of English-speaking children between the ages of 7 and 13 years recruited from 49 classrooms in primary and secondary schools in the East, North East, and South East of England. There were 1,135 children aged between 7 and 13 years (M_{age} = 10.22 years, SD = 1.45). Five hundred sixty nine children identified as girls and 563 children identified as boys (3 children did not wish to label themselves as boys or girls). Approximately 10% (N = 102) had a statement of special educational needs and 15.6% (N = 177) spoke languages in addition to English at home. Individual data on ethnicity and socioeconomic status were not available. Participants were drawn from 17 schools varying in socioeconomic and ethnic diversity. The median percentage of pupils eligible for free school meals (based on caregivers receiving state income support) was 14.4% (range: 0%–38%) and the median percentage of White British pupils was 57.7% (range: 6.5%–99.6%; Department For Education, 2019).

Sample 2 included English-speaking children between the ages of 8 and 13 years recruited from 37 classrooms in state-funded primary and secondary schools in the East and West Midlands of England. Of the 1,100 children enrolled in participating classrooms, 31 children were excluded because their caregivers did not provide consent for their participation and/or the children were unable to participate in the study unaided by a classroom assistant. A further 49 children declined to participate in the study. Of the remaining 1,020 children (93% participation rate) included in the study, 890 (87.3%) children participated in both study visits and 130 (12.7%) children participated in one study visit. Teachers completed questionnaires for 786 (77.1%) of the children. Missing teacher questionnaires were attributed to the cessation of testing due to the Covid-19 pandemic restrictions in March 2020.

Of 1,020 children in Sample 2, 556 identified as girls and 453 identified as boys. Children were aged between 8.27 and 13.27 years (M_{age} = 10.36, SD = 1.27). There were 475 8- to 9-year-old children, M_{age} = 9.27 years, SD = 0.45, 391 10- to 11-year-old children, M_{age} = 10.82 years, SD = 0.55, and 154 12- to 13-year-old children, M_{age} = 12.59 years, SD = 0.37. Participants were socioeconomically diverse: 23.2% (of 770 children) were eligible for free school meals (based on carers receiving state income support) and 28.9% (of 772 children) spoke languages in addition to English. The sample was ethnically diverse (based on data from 730 children): 51.5% White, 31.5% Asian, 8.1% Black, 6% Mixed Race, and 2.9% "other." One fifth of the children (18.2% of 768 children) had a statement of special educational needs.

Procedure

This study was preregistered on the open science framework (OSF; https://osf.io/rxyfh). The University of Birmingham Research Ethics Committee approved the study. Children participated in 60–90 min whole-class sessions. Sample 2 teachers completed ratings of children's social competence. Two research assistants led each testing session. The children completed all tasks individually on a school computer through PsyToolkit (Stoet, 2017). Sample 1 children completed the same tasks but entered their responses into article booklets. Responses from Sample 1 were digitally transcribed verbatim (including spelling errors). Teachers were present throughout but unaware of how each child performed as children recorded their answers in silence. High quality data were obtained by instructing children to work in silence during the session, monitoring to prevent conferring, and pacing activities using passwords so that children could not move on without instruction.

Measures

Theory of Mind

In the *Silent Film task* (Devine & Hughes, 2013) children watched five short film clips from a classic silent comedy depicting instances of deception, misunderstanding, and false belief. Children responded to a single question about each clip (read aloud by the research assistant), which required an explanation of a character's behavior. The research assistant did not play the next clip until all children had recorded an answer. Children's open-ended responses were later scored by two trained research assistants. Children received 2 points for accurate mentalizing given the context, 1 point for partially correct responses, and 0 points for inaccurate or irrelevant responses (see Devine & Hughes, 2016, for details on coding). The test and scoring manual are available at the OSF (https://osf.io/8x73r/).

In the *Strange Stories task* (Happé, 1994), the researcher read aloud five short vignettes, involving deception, false belief, and double bluff. The stories were displayed on a large screen for the children to see. Children answered an open-ended question about the characters' behavior. The researcher showed the next story when all children had recorded their response. Two trained coders later scored these responses. Correct responses involving accurate mentalizing received 2 points, partially correct responses received 1 point, and inaccurate or irrelevant responses scored 0 points (see White et al., 2009, for details on coding).

Following face-to-face training using examples and feedback, two graduate research assistants completed an unseen reliability testing set comprised data from 30 participants to each of the five strange stories and six silent film task items (i.e., 330 question–answer pairs). Interrater reliability (Krippendorf's α) ranged from .85 to .1.00 for the strange stories items and .87 to 1.00 for the silent

film task items. Having established interrater reliability each response was scored by one research assistant only. Research assistants met weekly with the study lead to discuss challenging cases and agree on consensus scores.

Verbal Ability

Children completed the multiple choice section of the *Mill Hill Vocabulary Scale* (Rust, 2008) to measure verbal ability. This test measures receptive vocabulary in 7- to 18-year-old children in group settings. Children selected a synonym for a target word from six possible response options and received 1 point for each correctly identified word. Items were summed and age residualized to create a verbal ability score.

Social Competence

Teachers of the children in Sample 2 completed the *Peer Social Maturity Scale* (Peterson et al., 2007). The scale captures peer-oriented social behaviors independently of age by asking teachers to rate children relative to their same-age peers in eight domains: making social overtures, assertion, leadership, sociable play, coping with peers, understanding others' needs, reading between the lines, and awareness of social situations. Teachers children using a 7-point scale ranging from "*very much less mature*" to "*very much more mature*" across seven items (Fink et al., 2013). High scores (7) indicated mature peer social interaction skills and low scores (1) indicated immature social interaction skills. Scores are associated with longer teacher-rated measures of social competence and peer-nominated social acceptance (Fink et al., 2013). Confirmatory factor analysis (CFA) using a robust maximum likelihood estimator showed that a one factor solution provided an initially poor fit to the data, $\chi^2(20) = 274.69$, root-mean-square error of approximation (RMSEA) = 0.128, comparative fit index (CFI) = 0.881, Tucker–Lewis Index (TLI) = 0.833. Inspection of modification indices revealed that the residuals for "assertion" and "leadership" were correlated. This improved model fit, $\chi^2(19) = 105.191$, RMSEA = 0.077, CFI = 0.961, TLI = 0.941, supporting the one factor structure. Item scores were averaged to create a social competence score.

Automated Scoring Systems

We compared three algorithms (i.e., SVM, BiLSTM, and DistilBERT) to create the automatic scoring system for the open-ended responses (Kovatchev et al., 2020). We trained the model to rate children's responses to each question from the Strange Stories and Silent Film tasks on a 3 point scale (i.e., 2, 1, and 0). We used Sample 1 for training and Sample 2 for testing the scoring systems.

We compared machine learning (i.e., SVM) with two deep learning neural network systems (i.e., BiLSTM and DistilBERT). We implemented the SVM classifier using Python's machine learning library scikit-learn (Pedregosa et al., 2011) using a radial basis function kernel and a standard set of text features: frequency of characters, words, character bigrams and trigrams, word bigrams and trigrams, and part of speech tags for each word. We implemented BiLSTM using TensorFlow 2 (Abadi et al., 2016). The network consisted of an input layer, an embedding layer, a bidirectional long short-term memory layer, a dense layer with "relu" activation, and a softmax layer. We also included a dropout layer for regularization. The embedding layer was initialized with random noise at the beginning of the training and trained for 20 epochs.

We implemented a standard Transformer classifier using KTrain Python library (Maiya, 2020). We used the pretrained DistilBERT model for initialization as available in the HuggingFace Python library (Wolf et al., 2020). We trained the network using the onecycle learning rate policy for four epochs. For both deep learning neural network systems we set the maximum length of the input to 35 tokens and used padding and truncating. We choose the maximum length so that it covered 99% of the children's responses. We tokenized each response using the Natural Language Toolkit Python library (Bird et al., 2009). We created a Python program based on the best-performing system to score responses to the strange stories and the silent film task and this is available at the OSF (https://osf.io/8x73r/).

Results

Analytic Approach

We carried out descriptive analyses in Jamovi Version 1.6 (The Jamovi Project, 2021) and used *Mplus* Version 8 (Muthèn & Muthèn, 2017) for latent variable modeling. Given item-level scores on the Strange Stories and Silent Film tasks were ordered categories with a limited number of responses, we used a mean- and variance-adjusted weighted least squares estimator (WLSMV; Yang & Green, 2011). We evaluated model fit using three standard criteria: a RMSEA of <.08, a CFI of >.90, and a TLI of >.90 (Brown, 2015). Table 1 and Table S2 show the extent of missing data for Sample 1 and 2. Missing data on the Silent Film and Strange Stories tasks did not exceed 6% on any item. Little's missing completely at random test was not significant in either Sample 1, $\chi^2(415) = 458.74$, $p = .068$, or in Sample 2, $\chi^2(364) = 357.73$, $p = .583$. Since data were missing at random, we handled missing data by using all available observations when estimating models with the WLSMV estimator (Asparouhov & Muthèn, 2010; Lei & Shiverdecker, 2020).

Descriptive Statistics

Table 1 shows descriptive statistics for Sample 1 and Sample 2. Table S2 shows the numbers of children achieving correct, partial and fail scores on each item of the strange stories and silent film tasks in Sample 1 and Sample 2. Table S3 shows polychoric correlations between test items in each sample. Figure S1 shows the distribution of summed scores on the strange stories and silent film tasks by age-group in Sample 1 and Sample 2.

Latent Factor Structure of the Silent Film and Strange Stories Tasks

We examined the factor structure of the strange stories and silent film task manual ratings by testing competing models in Sample 1 and cross-validating in Sample 2. Since the Akaike Information Criterion and Bayesian Information Criterion cannot be calculated from WLSMV models (Brown, 2015), we selected the best fitting model by examining the acceptability of overall model fit using the standard cutoff criteria noted earlier. We compared model fit for one- and two-factor models using robust χ^2 difference testing in *Mplus* (Muthèn & Muthèn, 2017). We examined whether factor variances and factor loadings were different from zero (Brown, 2015). In models with

Table 1
Descriptive Statistics for Sample 1 and Sample 2

Variable	Sample 1				Sample 2					
	M	*SD*	Range	*N*	*M*	*SD*	Range	*N*	*t*(*df*)	*p*
Age (years)	10.22	1.45	7.25–13.53	1,122	10.37	1.27	8.27–13.27	1,020	2.52 (2,140)	.012
Verbal ability	13.09	3.81	0–20	1,126	12.75	3.77	0–20	871	1.94 (1995)	.053
Strange Stories	5.34	1.94	0–10	1,131	5.65	1.95	0–10	980	3.63 (2,109)	.001
Silent Film	5.83	2.49	0–12	1,133	6.09	2.56	0–12	977	2.37 (2,108)	.018
Strange Stories (length)	10.09	4.05	2.33–29.60	1,130	14.78	8.02	1–51.80	980	17.29 (2,108)	.001
Silent Film (length)	9.88	3.65	1–26.83	1,133	14.92	7.42	2–51.67	977	20.22 (2,108)	.001
Strange Stories (mental)	—	—	—	—	4.04	1.02	0–5	980	—	—
Silent Film (mental)	—	—	—	—	4.09	1.31	0–6	977	—	—
Social competence	—	—	—	—	4.17	1.24	1–7	773	—	—

Note. Length = average response length in words; mental = responses referring to others' mental states regardless of context.

more than one latent factor, we examined the strength of factor correlations. Factors with correlations exceeding .80 were considered to lack of discriminant validity (Brown, 2015).

In Sample 1, we compared a two factor model, in which each indicator loaded onto either two latent factors (i.e., silent film task latent factor and strange stories task latent factor) and a one factor model. The factor metric was established by fixing the first factor loading to one. We tested a bifactor model where all items loaded onto a single ToM latent factor and items from the Strange Stories also loaded onto a task-specific factor (Geiser & Lockhart, 2012). Model fit indices are presented in Table 2. The two-factor model (Model 1) provided an acceptable fit to the data but modification indices suggested that residuals for Item 3 and Item 5 of the silent film task were correlated. Correlated residuals may reflect the fact that both clips involve one character not seeing something that the viewer sees and have been identified in previous studies using this measure (Devine & Hughes, 2016). A two factor model (Model 1 A), incorporating the correlated residuals, provided a good fit to the data. The latent factors were strongly correlated, *Std. Est.* = .87.

A one-factor model provided a good fit to the data (Model 2). Model comparison using robust χ^2 revealed no difference between the one- and two-factor models, $\chi^2(1) = 3.219$, $p = .07$. Since the latent factor correlation in the two-factor model exceeded .80, we selected the one factor model. The bifactor model provided a good fit to the data (Model 3), but factor loadings for the task-specific factor were nonsignificant, indicating that this factor did not account for any additional variance over the ToM latent factor.

As data were drawn from children in 49 classrooms, we accounted for nonindependence when computing standard errors and fit statistics by using the "complex" analysis command in *Mplus* with classroom as a cluster variable (Muthèn & Muthèn, 2017). This one-factor model provided a good fit to the data (Model 4). Given that each indicator consisted of a single item (rather than a composite score), completely standardized loadings of .30 were considered salient factor loadings (Brown, 2015). The factor loadings for 9 out of the 11 indicators exceeded this cutoff of .30 indicating that the latent factor was moderately-to-strongly associated with performance on each indicator (Brown, 2015; Gignac & Szodorai, 2016). The mean standardized factor loading was .38 and the median was .37. Loadings ranged from .18 (SF2) to .38 (SF4), all *p*s < .0001, for the silent film task and from .36 (SS1) to .62 (SS5), all *p*s <.0001 for the strange stories task. Factor loadings were consistent with those reported in studies using different measures of ToM in school-aged children (e.g., Osterhaus et al., 2016) and may reflect the heterogeneity of items (e.g., context, goal, mentalizing skills required; Table S1). The nonlinear model-based omega coefficient, $\omega_{u-\text{cat}} = .57$, for this one factor model estimated in *lavaan* (Rosseel, 2012) indicated that the correlation between the latent factor scores and total summed categorical item scores for the Silent Film and Strange Stories tasks was .75 (Flora, 2020; Yang & Green, 2015).

Table 2
Model Fit Indices for Sample 1 and Sample 2 (Based on Manual Ratings)

Model	Sample	χ^2	*df*	*p*	CFI	TLI	RMSEA	90% CI
Model 1. Two factor model.	1	64.05	43	.02	0.974	0.967	0.021	[0.008, 0.031]
Model 1A. Two factor model with correlated residuals for SF task.	1	44.60	42	.36	0.997	0.996	0.007	[0, 0.022]
Model 2. One factor model with correlated residuals for SF task.	1	48.29	43	.27	0.994	0.992	0.010	[0, 0.023]
Model 3. Bifactor Model with correlated residuals for SF task.	1	42.95	38	.27	0.994	0.991	0.011	[0, 0.024]
Model 4. One factor model with correlated residuals for SF task and clustering.	1	45.43	43	.37	0.996	0.995	0.007	[0, 0.022]
Model 5. One factor model with correlated residuals for SF task and clustering.	2	66.80	43	.01	0.932	0.914	0.024	[0.011, 0.034]

Note. CFI = comparative fit index; TLI = Tucker–Lewis Index; RMSEA = root-mean-square error of approximation; CI = confidence interval; SF = silent film task.

To cross-validate the one-factor model, we tested the model in Sample 2 (Model 5). Children were drawn from 37 classrooms so we accounted for nonindependence by computing standard errors and model fit statistics using the "complex" analysis command in *Mplus* with classroom as a cluster variable. Despite differences in sample characteristics (Table 1), the one-factor model provided an acceptable fit to the data and all items loaded significantly on the single latent factor (Model 5). The mean standardized factor loading was .36 and the median was .35. Factor loadings ranged from .20 (SF5) to .36 (SF4), all *p*s < .0001, for the Silent Film task and from .41 (SS2) to .52 (SS3), all *p*s < .0001, for the Strange Stories task. Loadings for 9 out of 11 indicators exceeded .30.

To rule out the possibility that the latent factor simply captured children's verbosity, we extended the one-factor model by regressing each item onto a variable measuring children's mean length of response (in words) to the items on the Silent Film task and items on the Strange Stories task. This model provided a good fit to the data, $\chi^2(54) = 70.646$, $p = .0132$, RMSEA = 0.022, 90% CI [0.01, 0.03], CFI = 0.967, TLI = 0.953. Standardized factor loadings remained significant for the Silent Film task items and Strange Stories task items, ranging from .25 to .45, all *p*s < .01, even when response length was considered.

To rule out the possibility that the latent factor captured children's propensity to use mental state language when describing others, we tested an alternative coding scheme by categorizing each response based on whether or not it referred to a character's mental states regardless of whether the reference was contextually appropriate. For example, a participant might use mental terms but do so in a way that is not appropriate for a given scenario (e.g., "He *wanted* to punish him"). Interrater reliability (based on 180 responses) identifying mental references was acceptable, nominal α = .80 (91.67% agreement). Supporting the original coding scheme, neither a one-factor model, $\chi^2(44) = 74.97$, $p = .003$, RMSEA = 0.027, 90% CI [0.016, 0.037], CFI = 0.819, TLI = 0.774, or two-factor model, $\chi^2(43) = 66.47$, $p = .012$, RMSEA = 0.024, 90% CI [0.011, 0.034], CFI = 0.863, TLI = 0.825, based on using mental-state words provided a good fit to the data.

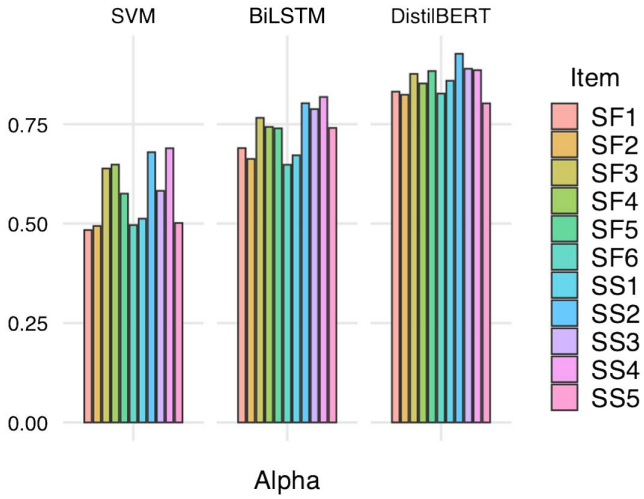
Reliability of Automated Scoring System for Advanced ToM

The automated scoring system was trained using data from Sample 1. We calculated interrater reliability (Figure 1) between trained manual coders and each automated scoring system using ordinal α (Hayes & Krippendorff, 2007). Table S4 shows item level scores for manual and automated ratings. Ratings based on SVM performed the worst, mean α = .57, range: .48–.69. Ratings based on BiLSTM performed better than SVM, mean α = .73, range: .65–.82, but some items fell short of acceptable levels of agreement. Ratings based on DistilBERT (deep learning algorithm) performed the best, mean α = .86, range: .80–.93.

Next, we tested whether latent factors based on manual ratings and the DistilBERT deep learning algorithm automated scores exhibited configural (i.e., same factor structure), metric (i.e., equal factor loadings), and scalar invariance (i.e., equal item thresholds; Brown, 2015). We implemented multiple-groups confirmatory factor analysis (MGCFA) by setting the scoring method (i.e., manual, DistilBERT) as a grouping factor and clustering each rating within participants by using the "complex" analysis command in *Mplus*

Figure 1

Estimates of Interrater Reliability of Automated Scoring Systems With Manual Ratings by Test Item (Krippendorff's α)



Note. SF = silent film task; SS = strange stories task; SVM = support vector machine; BiLSTM = bidirectional long-term short-term memory; DistilBERT = "distilled" Bidirectional Encoder Representation from Transformers. See the online article for the color version of this figure.

with participant as a cluster variable. We adopted this approach over within-person measurement invariance testing (Liu et al., 2017) because close agreement between manual and DistilBERT ratings produced polychoric correlations >.95 for some items. MGCFA with clustering by participant allowed us to account for nonindependence (i.e., each response was scored both manually and by the DistilBERT automated scoring system) while also testing for configural, metric, and scalar invariance (Narad et al., 2015; Walker & Göçer Şahin, 2020). We used a significant χ^2 difference test ($p < .05$) in conjunction with a decrease in CFI of >.002 to identify differences in model fit when comparing the nested models (Meade et al., 2008).

The assumption of strict measurement invariance was supported (Table 3). Manual and deep learning automated scoring system ratings of children's responses to the Silent Film and Strange Stories tasks showed identical factor structure, equal factor loadings (i.e., the latent factors were defined similarly), and equal item thresholds (i.e., similar levels of underlying ToM ability were associated with obtaining ratings of 0, 1 or 2). Latent factor means and variances from the automated scoring system did not differ from manual ratings. These results show that the automated scoring system captured the same latent factor as manual ratings by trained coders (Walker & Göçer Şahin, 2020).

Validity of Automated Scoring System for Advanced ToM

To establish the validity of automated scoring system data for use in research, we first replicated previous studies by examining the correlates of ToM for manually and automatically scored data (i.e., scores from the DistilBERT deep learning algorithm). We then extended previous work by testing the association between ToM

Table 3
Model Comparisons Showing That Manual Scores and Automated Scores Based on the Deep Learning Algorithm Exhibit Measurement Invariance

Model	χ^2	*df*	CFI	TLI	RMSEA	90% CI	Δ CFI	$\Delta\chi^2$	*df*	*p*
Equal factor structure	138.963	86	0.944	0.929	0.025	[0.017, 0.033]	—	—	—	
Equal factor loadings	142.739	96	0.951	0.944	0.022	[0.014, 0.030]	+0.007	7.926	10	0.636
Equal item thresholds	147.403	106	0.956	0.955	0.020	[0.011, 0.027]	+0.005	6.411	10	0.779
Equal factor variances	145.241	107	0.960	0.959	0.019	[0.010, 0.027]	+0.004	0.019	1	0.890
Equal factor means	148.632	108	0.957	0.956	0.020	[0.011, 0.027]	−0.003	2.413	1	0.120

Note. CFI = comparative fit index; TLI = Tucker–Lewis Index; RMSEA = root-mean-square error of approximation; CI = confidence interval.

latent factor scores and teacher-rated social competence for manually and automatically scored data. Table S5 shows the estimated correlations for manually and automatically scored data.

To examine age-related differences in ToM, we extended the MGCFA by regressing the ToM latent factor in the manually and automatically scored data onto age in years, verbal ability, and dummy indicators for gender (0 = girl, 1 = boy), special educational needs (0 = no, 1 = yes), and free school meal status (0 = does not receive free school meals, 1 = receives free school meals). Regression and covariance paths were estimated freely across groups. The model exhibited acceptable fit, $\chi^2(223) = 334.512$, RMSEA = 0.025, 90% CI [0.020, 0.030], CFI = 0.932, TLI = 0.919. Standardized parameter estimates are shown in Figure 2A. There were age-related increases in ToM latent factor scores (over-and-above differences in verbal ability) for both the manually scored data, *Std. Est.* = .33, 95% CI [.26, .41], $p < .0001$, and automatically scored data, *Std. Est.* = .35, 95% CI [.27, .42] $p < .0001$, such that older children performed better on the ToM latent factor than younger children even when verbal ability and demographic characteristics were considered. There was no difference between the strength of these two paths, $\chi^2(1) = 0.097$, $p = .755$.

In our second model, we extended the MGCFA by regressing teacher-rated social competence onto the ToM latent factor in the manually and automatically scored data. We regressed social competence onto potentially confounding variables: age in years, verbal ability, gender, special educational needs, and free school meal status. All covariates were permitted to correlate with each other and ToM. The model exhibited acceptable fit, $\chi^2(223) = 369.435$, RMSEA = 0.025, 90% CI [0.021, 0.030], CFI = 0.941, TLI = 0.928. Standardized parameter estimates are shown in Figure 2B. ToM scores based on manual, *Std. Est.* = .28, 95% CI [.15, .41], $Z = 4.173$, $p < .0001$, and automated, *Std. Est.* = .30, 95% CI [.16, .42], $Z = 4.28$, $p < .0001$, ratings were uniquely positively associated with teacher-rated social competence over and above potentially confounding variables such as age, verbal ability, gender, special educational needs, and socioeconomic status. There was no difference between the strength of these two paths, $\chi^2(1) = 0.057$, $p = .811$.

Discussion

The overarching aim of the present study was to examine the reliability and validity of machine learning and deep learning automated scoring systems for rating open-ended responses to the Silent Film and Strange Stories tasks, with a view to assessing suitability for large-scale, reproducible research on ToM. A single factor explained children's performance on both tasks in a sample of 1,135 7- to 13-year-old children and a separate sample of 1,020 8- to 13-year-old children. A deep learning scoring system showed high levels of interrater reliability with trained coders. Manual and deep learning automated ratings of the Silent Film and Strange Stories tasks exhibited strict measurement invariance, indicating that the automated scoring system captured the same underlying ability as manual ratings. Supporting the validity of automated scoring, manual and automatic ratings of ToM exhibited equivalent moderate associations with social competence.

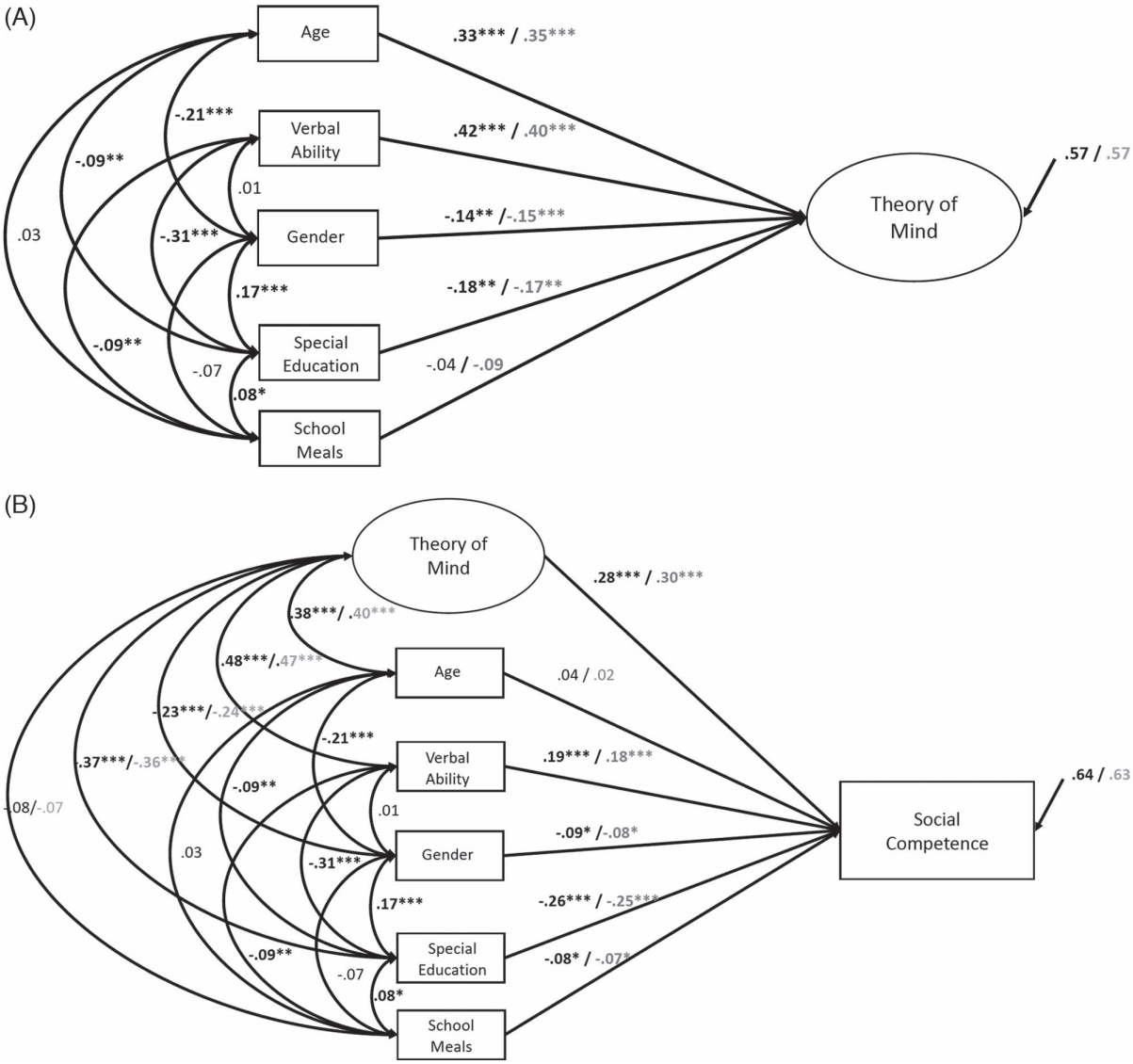
Individual and Developmental Differences

Our first aim was to replicate and extend previous research by testing and cross-validating the latent factor structure of the Silent Film and Strange Stories tasks across two large samples of children. Consist with prior work (e.g., Devine & Hughes, 2013), despite differences in stimuli and item content, a single latent factor explained performance on both tasks in middle childhood and early adolescence. Studies using CFA to examine the latent factor structure of ToM tasks remain relatively rare. Our results add to the growing literature on the nature of individual differences in advanced ToM (e.g., Devine, 2021; Osterhaus & Bosacki, 2022; Weimer et al., 2021) and extend existing work by using two separate samples to test and cross-validate the latent factor model.

Latent variable analysis can provide insight into the nature of age-related and individual differences in ToM. Our results replicated previous work showing age-related gains in ToM and, more specifically, on the Silent Film and Strange Stories tasks (e.g., Devine & Hughes, 2016). Age-related differences were not explained by verbal ability or a range of demographic characteristics. Moreover, individual differences in task performance were not explained by response length or children's propensity to use mental state words. These results suggest that the Strange Stories and Silent Film tasks capture differences in children's ability to mentalize taking context into account. Age-related differences in ToM may therefore reflect children's developing ability to apply insights about the mind to a range of situations taking context into account (Lagattuta & Kramer, 2021a, 2021b).

Responses to the Strange Stories and Silent Film tasks were underpinned by a common latent factor and scores on this latent factor accounted for approximately 57% of the variance in observed summed scores (Yang & Green, 2015). While the model provided a good fit to the data, the weak-to-moderate strength of some loadings suggest that individual items are unlikely to be

Figure 2
Correlates of Theory of Mind



Note. Standardized parameter estimates for correlates of theory of mind (Panel A) and association between theory of mind and social competence for manual ratings (Black) and automated ratings from the deep learning algorithm (gray; Panel B).
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

reliable indicators of this latent factor if used in isolation. Correlated residual terms might also suggest that performance on some items was driven by additional factors. Echoing these results, previous work using these tasks has shown that latent factor scores were highly stable (>.80) over a 1-month period whereas item level scores were less stable (Devine & Hughes, 2016). Researchers should therefore adopt latent factor scores (rather than summed scores or individual item scores) when using the silent film and strange stories tasks. Future research incorporating new test items to capture skills represented by existing items of the Silent Film and Strange Stories tasks (e.g., infer target's emotions based on beliefs *vs.* infer target's beliefs about another character's beliefs) but using different contexts (e.g., characters, situations) will elucidate whether individual differences in advanced ToM are driven by one or many factors.

Reliability of Automated Scoring of ToM

Our second aim was to examine the reliability of machine learning and deep learning automated scoring systems for the Silent Film and Strange Stories tasks. Measurement of item-level interrater reliability indicated that the deep learning automated scoring system yielded similar results as ratings by trained coders. Unlike machine learning (e.g., SVM algorithms) where features of the input text (e.g., frequencies of words, parts of speech etc.) are used to predict ratings (i.e., scores of 0, 1, or 2), deep learning uses an "end-to-end"

approach determining the relevant features of children's text responses. While this obviates the need to stipulate salient features in advance, it obscures those features of children's responses that provide the best predictors of children's ratings on each item. However, by using measurement invariance testing, we found that the automated scoring system for the Silent Film and Strange Stories tasks measured the same underlying latent ability as that captured by trained coders, suggesting that the scoring system treated children's responses in a similar way to trained human coders (Narad et al., 2015). The absence of any differential item functioning meant that the automated scoring system was no more biased than trained coders. These results suggest that automated scoring using deep learning provides reliable and unbiased estimates of children's ToM performance.

Beyond using deep learning algorithms for automated scoring of open-ended responses to the Silent Film and Strange Stories tasks, our study provides proof-of-principle that deep learning can be harnessed for scoring open-ended developmental assessments. The need for scalable and reproducible measures in developmental research is clear. Large samples are required because links between developmental measures and key social outcomes are typically modest in magnitude (e.g., Imuta et al., 2016). The replication crisis has highlighted how variation in task administration and scoring can alter results across studies (e.g., Poulin-Dubois et al., 2018). Machine learning and deep learning have been applied in psychological assessment across a range of domains (Takano et al., 2018; Victor et al., 2019) but, to our knowledge, deep learning algorithms have not been applied to developmental assessments. Many developmental measures rely open-ended responses (e.g., Livingston et al., 2021; Sher-Censor, 2015) making large-scale studies impractical and challenging to replicate (Iliev et al., 2015). Deep learning provides an innovative solution to this problem as it can be trained using multimodal data (Victor et al., 2019) including video, audio, and transcripts where expert ratings have already been completed making it possible to train automated scoring systems using existing scored archival data.

Validity of Automated Scoring of ToM

Our third aim was to test the validity of scores derived from the automated scoring system. Automated scoring system ratings of the Silent Film and Strange Stories tasks correlated with age and social competence to a similar degree as manual ratings suggesting that the scoring system generated good quality data for research. A unique association between performance on the Silent Film and Strange Stories tasks and teacher-rated social competence bolsters the view that these tests do not simply reflect differences in basic cognitive and linguistic abilities (Apperly, 2012; Hughes & Devine, 2015). By conventional standards (Cohen, 1988), the magnitude of the association between performance on the Silent Film and Strange Stories tasks and social competence was small to medium in strength (.28–.30). However, meta-analyses suggest that our results are consistent with effect sizes for associations between other widely studied developmental measures and "real world" outcomes (e.g., attachment security and social competence, $r = .19$, Groh et al., 2014). Moreover, recent meta-analyses suggest that the median effect size in psychological research is .19 and that .30 represents a large effect (Gignac & Szodorai, 2016).

Interestingly, the results reported here are larger in magnitude than those reported in meta-analyses linking ToM and aspects of children's social competence (e.g., Imuta et al., 2016) and support previous work showing that associations between performance on the Silent Film and Strange Stories tasks and social competence persist even when potential confounds are considered (e.g., Devine & Apperly, 2022). Showing incremental associations between performance on ToM tests is critical because social competence is shaped by other cognitive abilities (e.g., verbal ability, nonverbal ability) and contextual factors (e.g., familiarity, convention, routine; Lecce & Devine, 2021). It is estimated that the average sample size used in research on advanced ToM is 169 (Osterhaus & Bosacki, 2022). Given the magnitude of effect sizes reported here and elsewhere, large samples are necessary to understand the relations between ToM and social competence. The deep learning automated scoring system for the Silent Film and Strange Stories tasks can provide good quality data for future research aimed at isolating those aspects of social competence most strongly associated with ToM (Lecce & Devine, 2021). The Silent Film and Strange Stories tasks are efficient to administer (taking approximately 15 min in a group setting) and the availability of automated scoring enhances the scalability of these tasks. The testing materials and automated scoring system are available on the OSF (https://osf.io/8x73r/).

Caveats and Conclusions

Some limitations deserve note. First, despite evidence that scoring procedures for the Silent Film and Strange Stories tasks can be automated, data collection requires oversight by a trained researcher. Although, there are challenges with automated testing with children (e.g., engaging children's attention, ensuring children remain on task, minimizing conferring with caregivers or peers), the COVID-19 pandemic has created barriers to traditional in-person assessment in developmental and clinical research (Rhodes et al., 2020). Research is needed to ascertain whether automated ToM testing yields reliable and valid data without the need for researcher supervision.

Second, the Silent Film and Strange Stories tasks were developed in a European context and the automated scoring system has been trained entirely on English language responses from British children Aged 7–13 years. Although we tested an ethnically and socially diverse sample of more than 2000 children, further validation in different contexts is required. The Silent Film and Strange Stories tasks have been used in North America (McIntyre et al., 2018), Europe (Ronchi et al., 2020) and East Asia (Wang et al., 2016). Deep learning automated scoring systems are language agnostic. Models can be retrained by including different language examples. Multilingual transformer models (Devlin et al., 2019) can use existing English language examples in cross-lingual settings to improve performance on non-English data. Future studies will reveal whether the models can be trained to score non-English responses.

Notwithstanding these shortcomings, our study breaks new ground by harnessing deep learning algorithms to score children's open-ended responses to advanced ToM tests automatically. The automated scoring system for open-ended responses to the Silent Film and Strange Stories tasks can be used to generate scalable, reproducible, reliable and valid data about ToM in middle childhood and early adolescence. Beyond research on ToM, the study provides proof-of-principle that deep learning algorithms can be used to score open-ended psychological assessments in developmental research.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation* (pp. 265–283). https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

Alpaydin, E. (2016). *Machine learning: The new AI*. MIT Press.

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

Asparouhov, T., & Muthèn, B. O. (2010). *Plausible values for latent variables using Mplus*. http://www.statmodel.com/download/Plausible.pdf

Betz, N., Hoemann, K., & Barrett, L. F. (2019). Words are a context for mental inference. *Emotion*, *19*(8), 1463–1477. https://doi.org/10.1037/emo0000510

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. O'Reilly.

Bratsch-Hines, M. E., Carr, R., Zgourou, E., Vernon-Feagans, L., & Willoughby, M. (2020). Infant and toddler child-care quality and stability in relation to proximal and distal academic and social outcomes. *Child Development*, *91*(6), 1854–1864. https://doi.org/10.1111/cdev.13389

Brown, T. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Cassels, T. G., & Birch, S. A. J. (2014). Comparisons of an open-ended vs. forced-choice "mind reading" task: Implications for measuring perspective-taking and emotion recognition. *PLOS ONE*, *9*(12), Article e93653. https://doi.org/10.1371/journal.pone.0093653

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage*, *12*(3), 314–325. https://doi.org/10.1006/nimg.2000.0612

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, *21*(4), 458–474. https://doi.org/10.1037/met0000111

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

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

Department For Education. (2019). *Schools, pupils and their characteristics: Methodology document*. Crown Publications.

Devine, R. T. (2021). Individual differences in theory of mind in middle childhood and adolescence. In R. T. Devine & S. Lecce (Ed.), *Theory of mind in middle childhood and adolescence: Integrating multiple perspectives* (pp. 55–76). Routledge. 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. *Developmental Science*, *25*(1), Article e13137. https://doi.org/10.1111/desc.13137

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

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

Devine, R. T., White, N., Ensor, R., & Hughes, C. (2016). Theory of mind in middle childhood: Longitudinal associations with executive function and social competence. *Developmental Psychology*, *52*(5), 758–771. https://doi.org/10.1037/dev0000105

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-Training of deep bidirectional transformers for language understanding*. ArXiv. https://doi.org/10.48550/arXiv.1810.04805

Fink, E., Rosnay, M., Peterson, C., & Slaughter, V. (2013). Validation of the peer social maturity scale for assessing children's social skills. *Infant and Child Development*, *22*(5), 539–552. https://doi.org/10.1002/icd.1809

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

Geiser, C., & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state-trait analyses. *Psychological Methods*, *17*(2), 255–283. https://doi.org/10.1037/a0026977

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Girard, J. M., & Cohn, J. F. (2016). A primer on observational measurement. *Assessment*, *23*(4), 404–413. https://doi.org/10.1177/1073191116635807

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*(5–6), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Groh, A. M., Fearon, R. P., Bakermans-Kranenburg, M. J., van Ijzendoorn, M. H., Steele, R. D., & Roisman, G. I. (2014). The significance of attachment security for children's social competence with peers: A meta- analytic study. *Attachment & Human Development*, *16*(2), 103–136. https://doi.org/10.1080/14616734.2014.883636

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

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

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. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *41*(4), 483–490. https://doi.org/10.1111/1469-7610.00633

Hughes, C., & Devine, R. T. (2015). Individual differences in theory of mind from preschool to adolescence: Achievements and directions. *Child Development Perspectives*, *9*(3), 149–153. https://doi.org/10.1111/cdep.12124

Iliev, R., Dehghani, M., & Sagi, E. (2015). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, *7*(2), 265–290. https://doi.org/10.1017/langcog.2014.30

Imuta, K., Henry, J. D., Slaughter, V., Selcuk, B., & Ruffman, T. (2016). Theory of mind and prosocial behavior in childhood: A meta-analytic review. *Developmental Psychology*, *52*(8), 1192–1205. https://doi.org/10.1037/dev0000140

Kosinski, M. (2021). Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports*, *11*(1), Article 100. https://doi.org/10.1038/s41598-020-79310-1

Kovatchev, V., Marti, M. A., Salamo, M., & Beltran, J. (2019). A qualitative evaluation framework for paraphrase identification. *Proceedings of the international conference on recent advances in natural language processing* (pp. 568–577). INCOMA Ltd.

Kovatchev, V., Smith, P., Lee, M., Traynor, I. G., Aguilera, I. L., & Devine, R. T. (2020). *"What is on your mind?" Automated scoring of mindreading in childhood and early adolescence*. Proceedings of the 28th International Conference on Computational Linguistics (CoLING 2020), Barcelona,

Spain (pp. 6217–6228). https://doi.org/10.18653/v1/2020.coling-main.547

Lagattuta, K., & Kramer, H. J. (2021a). Advanced theory of mind in middle childhood and adulthood: Inferring mental states and emotions from life history. In R. T. Devine & S. Lecce (Eds.), *Theory of mind in middle childhood and adolescence: Integrating multiple perspectives* (pp. 15–36). Routledge. https://doi.org/10.4324/9780429326899-3

Lagattuta, K. H., & Kramer, H. J. (2021b). Advanced emotion understanding: Children's and adults' knowledge that minds generalize from prior emotional events. *Emotion*, *21*(1), 1–16. https://doi.org/10.1037/emo0000694

LaVoie, N., Parker, J., Legree, P. J., Ardison, S., & Kilcullen, R. N. (2020). Using latent semantic analysis to score short answer constructed responses: Automated scoring of the consequences test. *Educational and Psychological Measurement*, *80*(2), 399–414. https://doi.org/10.1177/0013164419860575

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

Lecce, S., & Devine, R. T. (2021). Social interaction in early and middle childhood: The role of theory of mind. In H. J. Ferguson & E. E. F. Bradford (Eds.), *The cognitive basis of social interaction across the lifespan* (pp. 46–68). Oxford University Press. https://doi.org/10.1093/oso/9780198843290.003.0003

Lei, P., & Shiverdecker, L. K. (2020). Performance of estimators for confirmatory factor analysis of ordinal variables with missing data. *Structural Equation Modeling*, *27*, 584–601. https://doi.org/10.1080/10705511.2019.1680292

Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486–506. https://doi.org/10.1037/met0000075

Livingston, L. A., Shah, P., White, S. J., & Happé, F. (2021). Further developing the Frith-Happé animations: A quicker, more objective, and web-based test of theory of mind for autistic and neurotypical adults. *Autism Research*, *14*(9), 1905–1912. https://doi.org/10.1002/aur.2575

Maiya, A. (2020). *ktrain: A Low-Code library for augmented machine learning*. ArXiv. https://doi.org/10.48550/arXiv.2004.10703

McIntyre, N. S., Oswald, T. M., Solari, E. J., Zajic, M. C., Lerro, L. E., Hughes, C., Devine, R. T., & Mundy, P. C. (2018). Social cognition and Reading comprehension in children and adolescents with autism spectrum disorders or typical development. *Research in Autism Spectrum Disorders*, *54*, 9–20. https://doi.org/10.1016/j.rasd.2018.06.004

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Meinhardt-Injac, B., Daum, M. M., & Meinhardt, G. (2020). Theory of mind development from adolescence to adulthood: Testing the two-component model. *British Journal of Developmental Psychology*, *38*(2), 289–303. https://doi.org/10.1111/bjdp.12320

Muthèn, L. K., & Muthèn, B. O. (2017). *Mplus: Statistical analysis with latent variables. User's guide* (8th ed.).

Narad, M. E., Garner, A. A., Peugh, J. L., Tamm, L., Antonini, T. N., Kingery, K. M., Simon, J. O., & Epstein, J. N. (2015). Parent-teacher agreement on ADHD symptoms across development. *Psychological Assessment*, *27*(1), 239–248. https://doi.org/10.1037/a0037864

Osterhaus, C., & Bosacki, S. (2022). Looking for a lighthouse: A systematic review of advanced theory-of-mind tests beyond preschool. *Developmental Review*, *64*, Article 101021. https://doi.org/10.1016/j.dr.2022.101021

Osterhaus, C., & Koerber, S. (2021). The development of advanced theory of mind in middle childhood: A longitudinal study from age 5 to 10 years. *Child Development*, *92*(5), 1872–1888. https://doi.org/10.1111/cdev.13627

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peterson, C. C., Slaughter, V. P., & Paynter, J. (2007). Social maturity and theory of mind in typically developing children and those on the autism spectrum. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *48*(12), 1243–1250. https://doi.org/10.1111/j.1469-7610.2007.01810.x

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

Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., Krist, H., Kulke, L., Liszkowski, U., Low, J., Perner, J., Powell, L., Priewasser, B., Rafetseder, E., & Ruffman, T. (2018). Do infants understand false beliefs? We don't know yet—A commentary on Baillargeon, Buttelmann and Southgate's commentary. *Cognitive Development*, *48*, 302–315. https://doi.org/10.1016/j.cogdev.2018.09.005

Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., Benitez, J., & Ocampo, J. D. (2020). Advancing developmental science via unmoderated remote research with children. *Journal of Cognition and Development*, *21*(4), 477–493. https://doi.org/10.1080/15248372.2020.1797751

Ronchi, L., Banerjee, R., & Lecce, S. (2020). Theory of mind and peer relationships: The role of social anxiety. *Social Development*, *29*(2), 478–493. https://doi.org/10.1111/sode.12417

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rust, J. (2008). *Raven's standard progressive matrices and Mill Hill vocabulary scale*. Pearson Education.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. ArXiv. http://arxiv.org/abs/1910.01108

Sher-Censor, E. (2015). Five minute speech sample in developmental research: A review. *Developmental Review*, *36*, 127–155. https://doi.org/10.1016/j.dr.2015.01.005

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

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Takano, K., Gutenbrunner, C., Martens, K., Salmon, K., & Raes, F. (2018). Computerized scoring algorithms for the Autobiographical Memory Test. *Psychological Assessment*, *30*(2), 259–273. https://doi.org/10.1037/pas0000472

Tamnes, C. K., Overbye, K., Ferschmann, L., Fjell, A. M., Walhovd, K. B., Blakemore, S.-J., & Dumontheil, I. (2018). Social perspective taking is associated with self-reported prosocial behavior and regional cortical thickness across adolescence. *Developmental Psychology*, *54*(9), 1745–1757. https://doi.org/10.1037/dev0000541

The Jamovi Project. (2021). *Jamovi* (Version 1.6) [Computer software]. http://www.jamovi.org

Victor, E., Aghajan, Z. M., Sewart, A. R., & Christian, R. (2019). Detecting depression using a framework combining deep multimodal neural networks with a purpose-built automated evaluation. *Psychological Assessment*, *31*(8), 1019–1027. https://doi.org/10.1037/pas0000724

Walker, C. M., & Göçer Şahin, S. (2020). Using differential item functioning to test for interrater reliability in constructed response items. *Educational and Psychological Measurement*, *80*(4), 808–820. https://doi.org/10.1177/0013164419899731

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

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. *Developmental Review*, *59*, Article 100945. https://doi.org/10.1016/j.dr.2020.100945

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684. http://www.jstor.org/stable/1132444. https://doi.org/10.1111/1467-8624.00304

White, S., Hill, E., Happé, F., & Frith, U. (2009). Revisiting the strange stories: Revealing mentalizing impairments in autism. *Child Development*, *80*(4), 1097–1117. https://doi.org/10.1111/j.1467-8624.2009.01319.x

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, *29*(4), 377–392. https://doi.org/10.1177/0734282911406668

Yang, Y., & Green, S. B. (2015). Evaluation of structural equation modeling estimates of reliability for scales with ordered categorical items. *Methodology*, *11*(1), 23–34. https://doi.org/10.1027/1614-2241/a000087

Received November 4, 2021
Revision received July 27, 2022
Accepted September 8, 2022 ■