Child Psychiatry & Human Development
https://doi.org/10.1007/s10578-023-01648-0

RESEARCH

Assessing Theory of Mind in Children: A Tablet-Based Adaptation of a Classic Picture Sequencing Task

Nicolas Petit[1,2] · **Ira Noveck**[3] · **Matias Baltazar**[1] · **Jérôme Prado**[2]

Accepted: 28 November 2023

Abstract

Correctly assessing children's theory of mind (TOM) is essential to clinical practice. Yet, most tasks heavily rely on language, which is an obstacle for several populations. Langdon and Coltheart's (Cognition 71(1):43–71, 1999) Picture Sequencing Task (PST), developed for research purposes, avoids this limitation through a minimally-verbal procedure. We thus developed a tablet adaptation of this task for individual application, engaging children's motivation and allowing response times collection. To assess this tablet-PST, we first tested a large sample of neurotypical children (6–11 years-old, N = 248), whose results confirmed the task's structural and content validity, and permitted the construction of three standardized clinical indices. In a second experiment, we applied those to previously diagnosed autistic children (N = 23), who were expected to show atypical TOM performance. Children's outcomes were consistent with what was hypothesized and confirmed the task's external validity and moderate clinical sensitivity. The tablet-PST thus appears as a suitable tool, providing detailed profiles to inform clinical decisions.

Keywords Theory of mind · Clinical assessment · Psychological tests · Psychometrics · Autism spectrum disorder · Neurodevelopmental disorders [7]

Theory of mind (TOM) refers to the ability to attribute mental states to others in order to explain and predict people's behaviors [1]. TOM is thought to be essential for human social interactions, grounding communication, self-representation, and moral reasoning [2]. Although TOM emerges in the first months of development [3], studies have found that TOM skills keep maturing up to late adolescence (e.g. Bosco et al. [4]). In the experimental literature, TOM is classically operationalized and assessed by measuring a person's ability to understand false beliefs and to distinguish such beliefs from one's own perception of the world [5]. In seminal false belief task, Wimmer and Perner [6] presented children with situations in which a character, who had left the scene knowing that a chocolate was in one location, did not see that it had been moved to another location. When 3-years-old are asked where the character would seek the chocolate, they tend to indicate the chocolate's actual position, rather than where the character *believes* the chocolate to be. This revealed an important milestone, because around the age of 4, children can be expected to understand false beliefs and the character's state of mind [7]. This paradigmatic situation was also used to assess the representation of second order mental states (Mary *believes* that John *thinks* that X), which are arguably more complicated to process and are mastered only later in development, i.e., around the age of 7 [8]. TOM then continues to mature during school-age (e.g., O'Hare et al. [9]). Beyond the paradigmatic understanding of false beliefs, the operationalization of TOM in assessment tools has widened to encompass the comprehension of different types of mental states, such as hidden intentions or affective states, and a variety of tasks exist for pre-school-age children [10]. TOM tasks targeting school-age, however, appear to display relatively low psychometric properties in neurotypical children [11].

Critically, TOM abilities are also predictive of interindividual differences that relate to children's social and cognitive skills. For example, early TOM skills can predict later

✉ Nicolas Petit
npetit.ortho@gmail.com

1 Le Vinatier Psychiatrie Universitaire, Lyon Métropole, France

2 Centre de Recherche en Neurosciences de Lyon (CRNL), Lyon, France

3 Laboratoire de Linguistique Formelle (LLF), Paris, France

Published online: 02 January 2024

social preference, friendship quality or academic achievement [12–14]. TOM impairments also affect quality of life in several clinical populations, such as schizophrenic patients [15], individuals with traumatic brain injuries [16], children with developmental language disorders [17] and children with autism spectrum disorder (from now on: *autism*). Therefore, in clinical practice, it is of obvious importance to properly assess TOM skills in order to detect a patient's difficulties and to guide interventions.

Autism is exemplary of a case that involves TOM difficulties, as these impairments have long been postulated to be central or primary to this condition [18]. This led to a large body of research that eventually questioned both the primacy and the universality of this marker while confirming the overall association between autism and TOM divergences, and ultimately rephrased it in terms of *differences* instead of a *deficit* (see Fletcher-Watson and Happé [19], but also Marocchini [20]). For example, Senju [21] proposed that autistic individuals[1] could pass standard false belief tasks when asked to, but that they would struggle to *spontaneously* use TOM to predict others' behavior. This reveals that identifying autistic-neurotypical differences could require subtle experimental designs (e.g., as opposed to simply observing mean accuracy on TOM tasks, see [22]). In line with this, Livingston and Happé [23] argued that autistic individuals might achieve the same level of performance as neurotypical participants in some highly-controlled lab-environments, but that they could rely on atypical processes that might be unusually effortful for them. This explains why some researchers employ response times (RT) to better characterize TOM mechanisms (e.g. Behrmann et al. [24]; Livingston et al. [25]). For example, Kaland et al. [26], who compared autistic to neurotypical adolescents, showed that mental state inferences generally take longer to process than physical state inferences, but that this slowdown was more pronounced in autism.

It is also important to note that most classic TOM tasks that are used in clinical practice largely rely on verbal material. This is the case for the standard false belief task [6], which uses verbal stories and verbal questions. However, it is also the case for other tasks that have been developed to assess more complex aspects of TOM (i.e., those developing later than 1st order false beliefs, see [11]). For example, the Strange Stories test [27], the Faux-pas test [28] or the Reading the Mind in the Eyes Test [29] are among the most used, but all either use verbally presented stories or rely on a verbal response format. This is an important confound, because language and TOM are known to be interrelated skills across development [30–32]. Such tasks are thus difficult to use in individuals with language disorders because failures can be attributed either to poor TOM abilities or to verbal difficulties. This is particularly relevant for autism, which is often found associated with comorbid conditions, especially language disorders [33].

Interestingly, TOM need not be assessed using verbal tasks. For example, Langdon and Coltheart [34] developed a picture sequencing task (*PST* from now on) precisely to measure TOM without using complex verbal material. In this task, participants are asked to order four picture-cards such that they form a coherent story. Each picture contains no linguistic material, or very basic 1-word or 2-words utterances (e.g. "Chocolate", "Hi", "Vroom", "Go away"). Critically, these stories can involve characters acting upon false beliefs and, as such, require TOM. Two other types of stories are used as controls (see Fig. 1): mechanical stories (involving physical causalities) and social-scripts (involving everyday social routines). The task includes four sequences of each type and each of the sequence is scored between 0 and 6. The PST has been used successfully to isolate TOM skills in a variety of typical or clinical populations [34, 35–38]. For instance, in Langdon and Coltheart's [34] study, the PST revealed that healthy adults with high schizotypal traits differed from low-schizotypal controls in the TOM condition only, but not in the control conditions. In neurotypical children, Rajkumar et al. [37] showed that this task could produce a TOM index which was independent from general intelligence but which correlated with another measure of social intelligence [37]. To our knowledge, however, no measurement of the psychometric characteristics of this task has ever been reported.

It is also worth mentioning that, so far, the PST has only been used in group designs, for example to compare average TOM skills in groups of patients versus neurotypical controls. The current work presents a computerized version of the PST (prepared for tablets) to be used as a clinical tool in order to assess the TOM skills of an individual child. Compared to paper-and-pencil solutions, assessments performed on computer tablets come with some advantages. First, they tend to be preferred by children [39] and favor engagement and attention. This is especially the case in autistic children [40], who may benefit from such interaction-free and predictable environments. Second, tablet-based tests also allow for perfectly standardized instruction, execution, feedback and scoring procedures, reducing the risk of experimenter bias or errors, while allowing for reliable group-testing [41]. Finally, tablets offer an easy and reliable solution to record

[1] Following Botha et al. ([67]), we use identity-first language to refer to autistic individuals, which is considered less offending and stigmatizing than person-first language ("person with autism") by the autistic community (e.g., [68]). Note, too, that the adjective "autistic" refers here to the whole autistic spectrum.

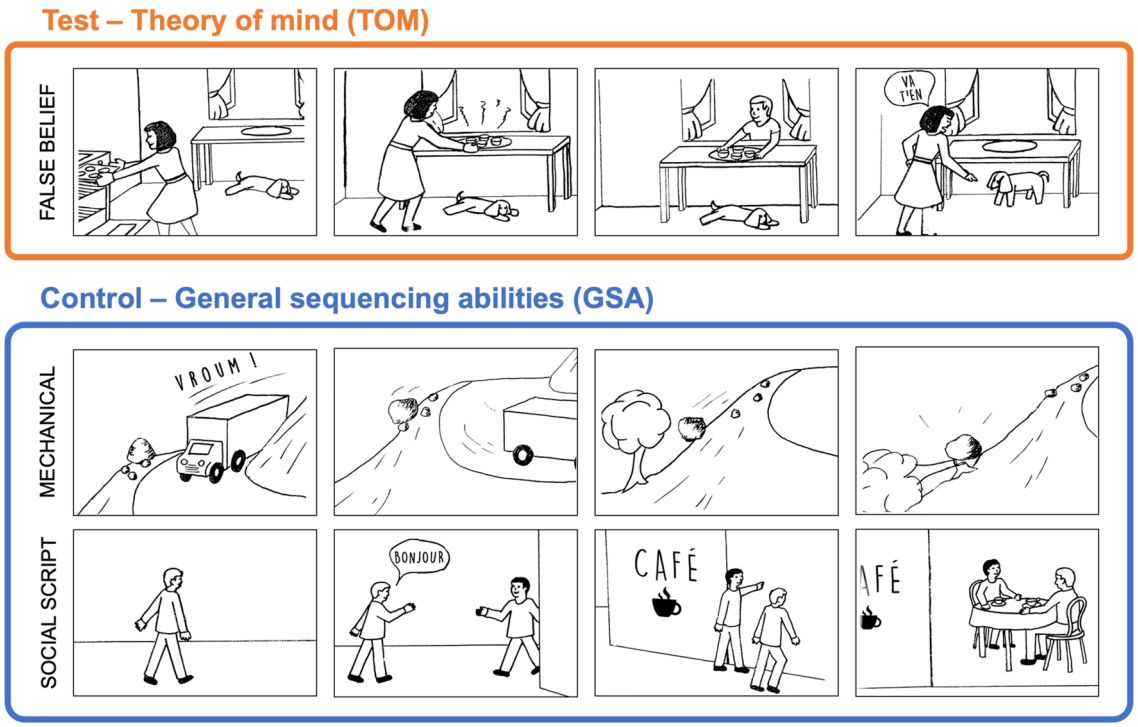

Fig. 1 PST stories example, involving false beliefs to assess theory of mind (above), or mechanical causalities and social scripts as control (below)

response times, which would provide relevant complementary information (besides just accuracy).[2]

In this study, we evaluated whether the tablet-PST could be used to assess individual children's theory of mind skills, as well as provided normative developmental data for the task. More specifically, we made four hypotheses regarding the characteristics of the tablet-PST. First, il should show good face validity as evidenced by a difference in difficulty between sequences that involve TOM and those that do not. Second, it should show a good structural validity, as evidenced by confirming a three-dimensions structure (with respect to its three types of sequences) along with a common ability to sequence pictures in general. Third, it should be sensitive to both (i) individual differences between children and (ii) group difference between neurotypical and autistic children. Specifically, the tablet-PST should discriminate between these two groups, although not perfectly since TOM is considered as divergent rather than absent in autism. Finally, consistent with prior research that used the PST [37], the tablet-PST should also demonstrate a good external validity within the group of autistic children by providing results that are consistent with a gold-standard TOM task. We did so by reporting two sets of results, a first experiment with neurotypical children, then a second experiment with a sample of autistic children.

Experiment 1

We began by testing a large sample of neurotypical children with the tablet-PST, and assessed its psychometric properties in terms of structural and content validity, as well as its sensitivity to interindividual differences. We focused on school-age children because they are in an appropriate developmental period, when TOM skills are maturing and can be the target of intervention (see [42], or [37] for data using the PST). School-age is also a critical period for the diagnosis of neurodevelopmental disorders (e.g., see Van 'T Hof et al. [43], for autism), yet when only few TOM assessments display good psychometric properties [11]. Assuming that our data meet psychometric standards, this would also represent normative data that could be used to derive standardized scores for individual applications.

[2] Interestingly, response times measures were included in Langdon & Coltheart's ([34]) original study but were later abandoned, after failing to reveal group difference between high vs. low schizotypal healthy adults. Given that response times have been shown to reveal meaningful effects in autism, we decided to include them on the tablet-PST.

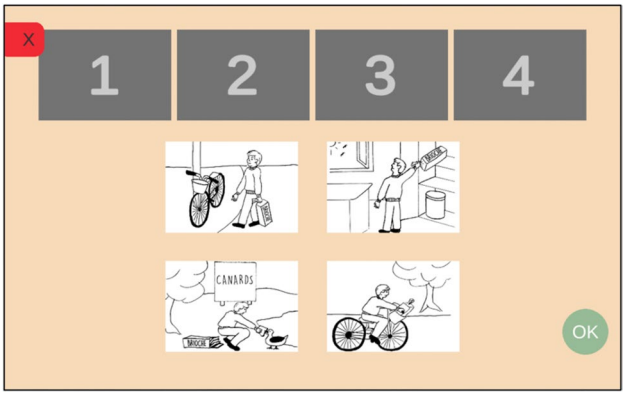

Fig. 2 Task set up for a practice item as it appears on the tablets of the participants

Methods

Participants

We recruited 248 children in primary schools for this experiment (mean age = 8.3 years old, min = 6.0, max = 11.0, 51% female). All participants were native French speakers. Fifty-two participants were in 1st grade, 58 in 2nd grade, 50 in 3rd grade, 42 in 4th grade and 46 in 5th grade. Parental information and consent were collected before testing. All parents declared that their child presented no learning or neurodevelopmental disorder (i.e., autism, attention deficit, or language-related disorder) and no sensory (visual or auditory) or motor disorders that would interfere with their ability to use a tablet. Parents also provided information regarding their socio-economic environment, through (1) the Family Affluence Scale (FAS, [44]), which is a good indicator of family wealth across countries, and (2) their education levels, which were coded on a scale from 0 (no diploma) to 7 (PhD). The FAS and the parental education level were standardized on a scale ranging from 0 to 100 (representing the range of minimum to maximum possible values) and averaged to provide a socio-economic status composite (SES; mean = 58.4; SD = 16.1). This study was part of a larger project on pragmatic inference skills across development, which received ethical approval from the local IRB (Comité de Protection des Personnes Sud-Est I, ID RCB 2019-A01721-56).

Material

As in the original task, the tablet-PST includes four trials of each type (mechanical, social scripts, false beliefs) as well as two training sequences. Each of these will be referred to as an *item*. The minimal verbal content of the original pictures (as depicted in Fig. 1) was translated into French. The task was implemented in an application designed for tablet computers. At the beginning of each trial, four pictures were presented in the lower part of the screen in a pseudo-random order (see Fig. 2) while four empty slots were located in the upper part of the screen. Children had to select and move the four pictures to the appropriate slots so that the sequence was in order and described a coherent story. They were allowed to modify the sequence as often as they wanted after moving the pictures if they changed their mind. When satisfied, they validated the sequence by pressing *OK* on the bottom right side of the screen. In accordance with the original procedure [34], accuracy on each item ranged from 0 to 6 based on the appropriate position of each picture (2 points for the first and the last picture, 1 point for the 2nd and the 3rd). Accuracy as well as response time (RT, i.e., time from the beginning of the trial to validation) was recorded by the device at the end of each trial. Trials were pseudo-randomized in a unique trial list. Feedback was automatically provided by the device for the 2 training sequences (a green screen in case of success, a grey screen in case of failure), which children had to perform successfully to move on to the rest of the task (these were repeated as many times as necessary).

Procedure

Assessment took place in children's usual classrooms, i.e. in a collective setting, with 15 to 25 children completing the task simultaneously. Each child was provided with a tablet. Automatic audio instructions were included in the task, asking children to move the four pictures to the appropriate slots so that the sequence was in order and described a coherent story, for each sequence. The task began with the two practice items, during which two experimenters remained available and could help children to correctly sequence pictures through verbal guidance. Then the task automatically continued with task items, during which no help was provided to children. Children performed the task at their own pace and were asked to remain quiet. During task completion, two experimenters made sure that children would not copy one other. The task, which lasted about 10 min, was performed after other experimental tasks were completed (these assessed pragmatic skills and are not reported here and took about 20 min).

Analysis

All analyses were conducted in R [45]. These involved three steps. First, we assessed the internal consistency and the factorial structure of the task. Internal consistency was assessed with Cronbach's alpha. The impact of each item on this value was also systematically measured. Factorial structure was assessed with a confirmatory factor analysis (CFA) using the *lavaan* package [46]. Specifically, we compared

a unifactorial structure to a theoretical three-dimensions structure.

Second, we used mixed effect models on single trials to analyze accuracy and response times at the group level. In these analyses, we were not interested in the differences among the control items (i.e., between the mechanical and social scripts), but in using them as controls for false-belief items. We thus followed the rationale of Langdon and Coltheart [34] (see also [37] for children) and contrasted TOM items to non-TOM items as a binary variable (*item-type*).[3] With respect to accuracy as dependent variable, the model included item-type, age, gender and SES as fixed effects, as well as the interactions of item-type with the latter three. The model with response times as dependent variable included these factors plus accuracy on each trial and its interaction with item-type as an additional fixed factor. All models included random intercepts for participants and items. Age was included as a continuous predictor (in years with decimals), while considering both a linear and a quadratic term (to capture potential non-linear effects). The response time analysis was run on log-transformed values to correct for skewness. Outliers were removed, first roughly at a group level by discarding trials with a response time shorter than 5 s or longer than 80 s, and secondly at a participant-specific level by excluding trials with a response time greater and lesser than 2.5 SD from each individual's mean RT. Models were fit with the *lme4* package [47], sum contrasts were used for all factors, and continuous predictors were centered on the mean. Effects were assessed with Satterthwaite's degrees of freedom from the *lmerTest* package [48], and post-hoc contrasts were computed with the *emmeans* package [49].

Third, for the sake of preparing the clinical application, we calculated summary measures for each participant and used it to fit linear regressions. These regressions were meant to allow for the comparison of an individual's actual performance to their predicted performance. The advantage of using regressions to calculate discrepancy between a single participant and a reference group is that, instead of considering arbitrarily-defined control sub-groups to provide normative data (for example age-groups or SES-groups), the whole sample size and its variability is exploited to improve the statistical power of standardized scores calculation [50]. Rather than specifying *a priori* the predictors to be included in those models, we used the factors that were shown to be significant at group-level analyses (step 2); for example, we specified age as predictor if and only if it was shown to be linked to performance. To anticipate, based on group-level results, (a) we selected age and SES as predictors of children's average accuracy on non-TOM items (as a General Sequencing Abilities index, GSA), (b) we used this GSA and both a linear and quadratic term for age to predict children average accuracy on TOM items, and (c) we used the mean response time to non-TOM items and the average accuracy on TOM items to predict the average response time to TOM-items.

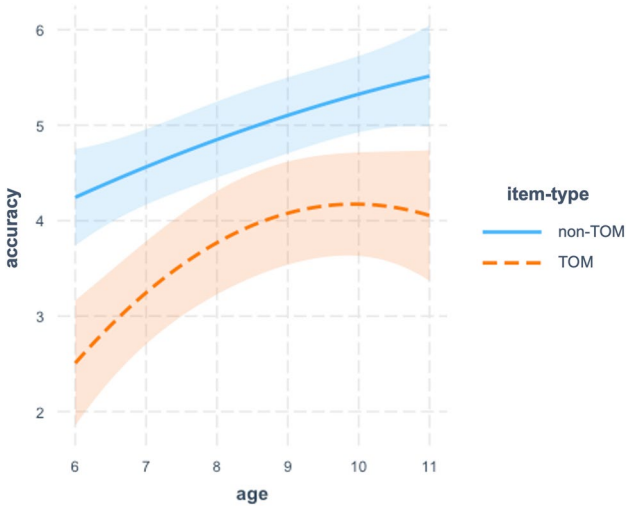

Fig. 3 model's predicted accuracy as a function of age and item-type (shaded areas represent 95% confidence intervals)

Results

Task Validation

Across all items, the overall Cronbach alpha was 0.75, suggesting acceptable internal consistency. Importantly, removing each item decreased this value (for details, see the supplementary materials, Table S.1.1). The CFA revealed that the three-factor structure (TOM, Mechanical items, Social-scripts) provided a good fit to the data ($CFI = 0.93$, $TLI = 0.90$, $RMSEA = 0.047$), which was significantly better than a unidimensional model ($X^2(3) = 17.8$, $p < .001$). The three factors covaried (standardized estimates varied from 0.63 to 0.90, all $ps < 0.001$, see Figure S1.1 in supplementary materials for detailed estimates).

Group Analysis

The linear mixed effect model with accuracy as dependent variable (Table 1) revealed a main effect of item-type, indicating that participants were less accurate on TOM items than the control (non-TOM) items. However, item-type interacted with age: while accuracy on non-TOM items increased linearly with age, accuracy on TOM items reached a plateau at around 9–10 years old (see Fig. 3). There was also a main

[3] To make sure that grouping together the two control conditions did not hide important effects, we also ran the analyses contrasting TOM, mechanical and social scripts, and report the results in supplementary materials.

Table 1 Estimated parameters of the model explaining accuracy

Parameter	Coefficient	Standard error	*p*
(Intercept)	3.14	0.26	< 0.001
item-type	0.95	0.39	**0.009**
gender	0.05	0.11	0.659
age	22.06	3.09	**< 0.001**
age^2	− 5.83	3.09	0.059
SES	0.02	0.00	**< 0.001**
item-type * gender	0.16	0.14	0.237
item-type * age	− 4.95	3.71	0.182
item-type * age^2	8.59	3.70	**0.020**
item-type * SES	0.00	0.00	0.445

Bold lines represent significant terms ($p < .05$)

age^2 = quadratic term for age

Table 2 Estimated parameters of the model explaining log-transformed response times

Parameter	Coefficient	Standard error	*p*
(Intercept)	9.81	0.08	< 0.001
item-type	0.07	0.11	0.553
accuracy	**0.01**	**0.00**	**0.004**
SES	0.00	0.00	0.126
age	**− 4.32**	**0.99**	**< 0.001**
age^2	**2.55**	**0.98**	**0.010**
gender	− 0.07	0.04	0.056
item-type * score	**0.03**	**0.01**	**< 0.001**
item-type * SES	0.00	0.00	0.175
item-type * age	**2.99**	**0.79**	**< 0.001**
item-type * age^2	0.59	0.77	0.446
item-type * gender	− 0.01	0.03	0.774

Bold lines represent significant terms ($p < .05$)

age^2 = quadratic term for age

effect of SES, indicating that accuracy generally increased with higher SES. No other main effect or interaction was significant.

The linear mixed model with response times as a dependent variable (see Table 2) revealed a significant interaction between item-type and accuracy (see Figure S1.2 in supplementary materials). While response times did not depend on accuracy for control items (ß = -0.01, $SE = 0.01$), it increased with accuracy on TOM items (ß = 0.03, $SE = 0.01$). There was also an interaction between item-type and age. Responses become faster overall with age, reaching a plateau at roughly 9 years of age. TOM items were associated with longer response times when compared to non-TOM items and this difference increased with age.

To make sure that the social script/mechanical items grouping did not hide important effects, we exploratorily reran these two models with 3 item-types instead, which revealed the very same patterns (see in Supplementary materials the detailed models outputs in Tables S1.2 and S1.3, and the predictions in figures S1.3 and S1.4).

Individual Indices

Summary measures were then calculated for each individual. Specifically, we calculated a TOM-accuracy score by averaging accuracy across TOM items, and a General Sequencing Abilities (GSA) score by averaging accuracy on control items (following [37]). Both indices showed satisfying distributions, avoiding floor or ceiling effects across age groups, although the GSA approached ceiling by 5th grade (see Figure S1.5 in the supplementary materials). Log-transformed response times were also averaged for these two item-types (TOM and non-TOM).

Based on the effects observed at the group-level (see above), three linear models were fitted to allow for the calculation of individual discrepancy indices (which will be useful for Experiment 2 below). First, we calculated a model explaining each child's GSA based on their age (ß = 0.26, $SE = 0.04$, $p < .001$) and SES (ß = 0.02, $SE = 0.004$, $p < .001$). This provides a control index of the extent to which a child is able to sequence pictures in general. Second, we calculated a model explaining each child's *TOM-Accuracy* based on their GSA (ß = 0.52, $SE = 0.07$, $p < .001$) and their age with a linear (ß = 1.94, $SE = 0.69$, $p < .001$) and a quadratic term (ß -0.0007, $SE = 0.00028$, $p = .01$). This allows one to characterize variations of accuracy on TOM items after regressing out children's age and ability to sequence pictures in general. Third, we calculated a model explaining each child's *TOM-Response times*, based on their accuracy (i.e., accounting for speed-accuracy tradeoff, ß = 0.057, $SE = 0.01$, $p < .001$) and their own baseline RT on control items (ß = 0.90, $SE = 0.05$, $p < .001$). This model allows one to estimate the effort associated specifically with TOM items.

Discussion

In this experiment, we tested a tablet implementation of the PST on 248 neurotypical school-aged children. The confirmatory factor analysis provided evidence in support of the structural validity of the tablet-PST, with 3 dimensions differentiating mechanical causalities, social scripts and false beliefs items. The non-negligible shared variance between the three factors (also visible in the Cronbach alpha demonstrating internal consistency) confirms the appropriateness of averaging control items together to be compared to TOM items, in line with previous uses of the PST [34,

37]. Even though we did not directly compare the results of the original paper-and-pencil PST to our tablet-PST in our sample, our results are in line with previous studies which used the original task with healthy participants, which typically observed that TOM items were harder than mechanical and social-scripts items. This is visible with respect to both accuracy [35–38] and response times [34]. It is interesting to note that children in our sample performed relatively well overall and better than children from a previous study by Rajkumar et al. [37], who also tested a group of neurotypical children (between the ages of 8 and 11). Although it might be argued that the difference comes from the tablet implementation of our task, this is not likely because tablet-devices have provided reliable results compared to paper-and-pencil tests elsewhere [39]. Another possibility, which could be investigated in future studies, is that the difference stems from cultural or educational differences between our setting (France) and theirs (Southern India). Critically, however, this difference is orthogonal to the contrast observed between TOM and control items. That is, our results remain in line with the TOM literature, which shows that TOM continues to develop during school age (e.g. O'Hare et al. [9]) and that intelligence in general can account for some, but not all variance on TOM performance [51]. This explains why TOM and control items are linked but follow slightly different developmental patterns.

Langdon and Coltheart [34], followed by Rajkumar et al. [37], previously used the PST to calculate a TOM composite by subtracting accuracy in TOM items from accuracy in control items. We followed this general approach for our clinical index, except that we determined the coefficient linking control to TOM accuracy on the basis of the group-level data. Moreover, we added age as a supplementary predictor, since we observed that both item-types followed a different developmental pattern. By doing so, our indices are based on theoretical considerations, but are also data-driven, i.e., they are based on the effects which emerged from the reference data. For example, the control index targeting how easy it is for a child to sequence pictures in general (GSA) was based on age and SES, since both factors appeared to play an important role at a group-level, regardless of item-type. Conversely, SES was not included in the TOM-Accuracy index because no interaction was observed between item-type and SES. Including GSA as a predictor of TOM accuracy should thus be enough to account for that variation. Using the same reasoning, gender was not taken into consideration to predict a child's performance. That is, while some studies suggested a TOM advantage for girls [52], this was not evident in our data, in line with Charman et al. [53]'s large study that observed this slight advantage, but only in early pre-school development.

Overall, this experiment provided outcomes that were consistent with prior findings in the literature, demonstrating the face validity of the tablet-PST. This also constituted normative data against which individual data could be compared. A next important step would be to find support for the task's utility by applying the clinical indices to a sample of children with a condition known to experience difficulties in theory of mind, such as autism. This is what we turn to in the next experiment.

Experiment 2

In this second experiment, we turned to a clinical population to pursue the validation of the tablet-PST for clinical use. Autistic children with normal language and intellectual skills appeared as an appropriate clinical condition for that purpose since, as described in introduction, they are expected to display TOM divergence compared to neurotypical children, but not an absence of TOM; the task should capture this by discriminating these two subgroups, although not categorically. Moreover, we expected this sample to display a correlation between the scores from the critical condition of the tablet-PST and those from a paradigmatic false-belief story task. In contrast, in order to show the specificity of the tablet-PST's ability to capture TOM, we also expected a much weaker (or even lack of) correlation with a parental report of Attention Deficit / Hyperactivity Disorder (ADHD) symptoms, which is a distinct psychological construct. This would demonstrate the external validity of the tablet-PST, both in a convergent and divergent way.

Methods

Participants

We recruited a sample of 23 autistic children (3 girls), mainly through the department of neurodevelopmental disorders at the hospital *Le Vinatier (Lyon, France)*. All children met the DSM-5 criteria for autism spectrum disorder and received their diagnosis from an experienced child psychiatrist, with no associated intellectual disability, language disorder or attention deficit disorder. Children's assessment results regarding their autism diagnosis (through the SCQ and/or ADOS-2), as well as their evaluation of intelligence were extracted from their medical records (see Table S2.1 in the supplementary materials), which confirmed the diagnosis. Regarding intellectual abilities, recent measures were not always available (data were not available for three children) and could not be averaged since they were collected with different instruments and gathered in different composites. However, a careful look at all individual results revealed that all

children fall in the range of what is considered normal to normal-superior intelligence (all IQ scores are in the 70–130 range, with most scores in the 100–130 range). All participants attended regular schools. Children were between 6;7 and 10;8 years of age (mean = 8.7 years old, *SD* = 1.36). SES measures were collected as they were in Experiment 1 and were found to be comparable to those of the control group (mean = 55.2, *SD* = 17.0, $t(25.8) = 1.40$, $p = .17$). Parental information and consent were collected before testing. This experiment was part of the same project as the first experiment, which received ethical approval from the local IRB (Comité de Protection des Personnes Sud-Est I, ID RCB 2019-A01721-56).

Materials

Children were tested with the tablet-PST, as in Experiment 1. To assess convergent validity, children were also tested with another standard 1st and 2nd order false belief task. This task, taken from the French battery EVALEO 6–15 [54], relies on verbal stories with picture support and verbal open questions targeting characters' (false) beliefs. Stories from this task are directly inspired by classical false belief stories from Wimmer and Perner [6] or Perner and Wimmer [8]. Five stories are presented, providing a total raw score ranging from 0 to 7. It showed an acceptable internal consistency in the large sample of children of the validation study (alpha = 0.77, [54]).

To assess divergent validity, parents also completed the French version of the ADHD-RS [55], a parental questionnaire which assesses inattention and impulsivity symptoms. The ADHD-RS contains 18 child behaviors, whose frequency has to be evaluated by parents. Its French version showed a good factorial structure and internal consistency [55]. Each question is scored from 0 to 3 and the total score varies between 0 and 54.

Procedure

Children were tested individually in a quiet office of the hospital. Efforts were made to let children progress through the task at their own pace (as in Experiment 1, we did not over-monitor what they were doing). Since this study was part of a larger research project, the tablet-PST was presented together with other cognitive tasks assessing pragmatic skills. Tasks were presented in an individualized order with as many pauses between tasks as needed in order to adapt to each child's level of attention and engagement, but the sequence of items within the tablet-PST did not differ from the one in Experiment 1.

Analysis

Our analyses involved three steps. First, we assessed convergent and divergent validity of the tablet-PST. Convergent validity was assessed by calculating the correlation between the mean score on TOM items and the EVALEO false-belief raw score. Divergent validity was assessed by calculating the correlation between the mean score on TOM items and the ADHR-RS questionnaire. Spearman rank correlations were used, given the sample size.

Second, we analyzed raw by-trial data at the group level with mixed effect models, by adding the autistic group to the control sample of Experiment 1. As the limited sample size of the autistic group inflated the risk of overfitting and convergence issues, we used the minimally needed models to observe the specific group effects that were of interest. We typically dropped SES which was not implicated in interactions with item-type in Experiment 1, and which did not differ across groups. For both models, we specified group, item-type and age (and their interactions) as fixed effects. We also included random intercepts for items and participants. We then used backward elimination and removed the terms that were not significant from this initial model, in order to reduce model complexity. The model was fitted on all data points for accuracy. For the analysis of response times, we analyzed log-transformed response time data after considering, in line with Langdon and Coltheart [34], successful responses only (score = 6) in order to reduce complexity in the model. We removed outliers that deviated by more than 2.5 SD in either direction from each group's mean (individually personalized thresholds were not possible due to the exclusion of several items, see Results). Models were fit with the *lme4* package [47], sum contrasts were used for all factors except group, where the neurotypical group was used as baseline, and continuous predictors were centered on the mean. Effects were assessed with Satterthwaite's degrees of freedom from the *lmerTest* package [48], and post-hoc contrasts were computed with the *emmeans* package [49]. The effects that did not involve group have already been described in detail in Experiment 1; for this reason and for clarity, we only reported the group effects and interactions that were of interest, while the complete results of the analysis are presented in the supplementary materials.

Third, we computed summary measures for each participant, according to the procedure of Experiment 1, and discrepancy indices based on the normative sample. That is, we used the linear regressions fitted in Experiment 1 to predict each participant's performance and compared it to their actual performance. To qualify this discrepancy, we followed Crawford & Garthwaithe's [50] methodology, which was shown to be more robust than using the basic model standard error, using the R script developed

by Arcara [56]. This discrepancy can be expressed as an estimated percentile rank that ranges from 0 to 100. For comparability, the RT index was reversed, so that a lower estimated percentile rank is always indicative of greater effort or difficulty. We then compared the two groups' standardized scores for the three indices with t-tests and through visual examinations of the data distributions.

Results

Task Validation

As expected, the tablet-PST TOM average score was correlated with the EVALEO false belief score (ρ = 0.49, $p = .02$), indicating convergent validity. Also, as expected, the tablet-PST TOM average score was not significantly correlated with the ADHR-RS total score (ρ = 0.16, $p = .50$), indicating divergent validity.

Group-Level Results

The linear mixed model explaining accuracy (for the complete output see table S2.2 in the supplementary materials) revealed no main effect of group (ß = -0.09, $SE = 0.21$, $p = .68$) but a group by item-type interaction (ß = 0.45, $SE = 0.23$, $p = .046$), such that the autistic group performed better than neurotypical children with respect to control items (ß = -0.15, $SE = 0.22$, $p = .50$) but worse with respect to TOM items (ß = 0.32, $SE = 0.26$, $p = .21$).

In terms of response times, the model was fitted on 1914 data points (after exclusion of 40% of non-perfect responses[4], then of 1.5% of outliers, see complete output in table S2.3 in the supplementary materials), revealing a main effect of group (ß = 0.24, $SE = 0.06$, $p < .001$), with autistic participants being slower overall, as well as a three-way group* item-type *age (ß = 0.15, $SE = 0.07$, $p < .05$) interaction. This interaction revealed that the group difference was not constant over age and item-type (Figure S2.1, supplementary materials). While TOM items took longer in general than control items, this difference appeared constant over age in the neurotypical group (ß = 0.01, $SE = 0.02$, $p > .05$) but tended to decrease in the autistic group (ß = -0.10, $SE = 0.05$, $p = .06$). We thus computed group-differences on this item-type contrast at minimum, mean and maximum age, which revealed that the slowdown associated to the TOM items tended to be larger in the autistic group than in the neurotypical group at 6 years-old ($p = .05$) but not at mean or at maximum ages ($ps > 0.05$).

[4] In line with Experiment 1, perfect responses (score = 6) were not distributed equally across item types: they represented 39% of responses to TOM items and 70% of the responses to non-TOM items.

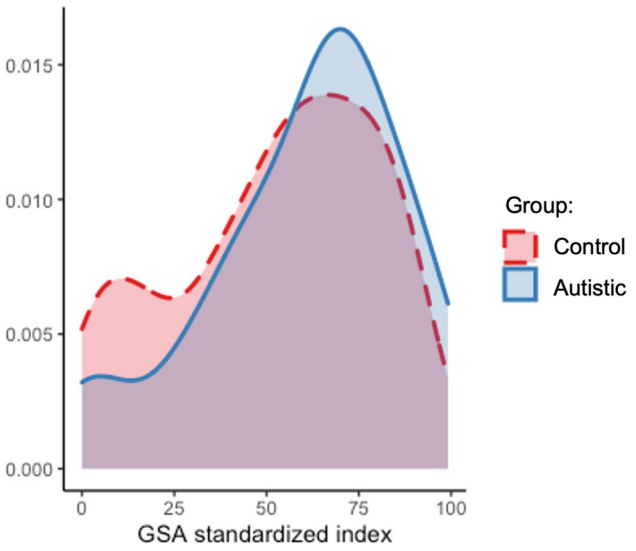

Fig. 4 GSA standardized index' distribution for the autistic and the control group

Individual Indices

For each child, we estimated their GSA, TOM-Accuracy and TOM-RT percentile ranks, based on the linear regressions fitted in Experiment 1 (note that the results from the control group in this analysis are retrieved from Experiment 1 so they are not new; they are nevertheless displayed for comparison with the autistic group). The Welch two-sample t-tests comparing the groups on those indices failed to reveal significant differences with respect to GSA ($t(26.5) = -1.32$, $p = .20$), TOM-Accuracy ($t(25.7) = 1.52$, $p = .14$) and even TOM-RT which did approach significance ($t(27.5) = 1.96$, $p = .06$). However, visual inspection of the distributions suggests interesting effects. On the one hand, it confirms the absence of group-difference for the GSA (see Fig. 4). On the other hand, it suggests slightly different distributions for the TOM-Accuracy and TOM-RT indices. Specifically, while the control group appears homogeneously distributed (as expected), a different pattern emerges for the autistic group. As shown in Fig. 5, joint distribution for autistic individuals peaks in the bottom left quarter, where accuracy is poorer and response-times are longer than expected. Although only one fourth of the control group is distributed in this quarter, half of the autistic sample can be found there ($X^2(1, N = 271) = 5.10$, $p < .05$).

Based on these results (and following Langdon & Coltheart's, [34] rationale) we computed a TOM-Effort composite, which is an average of the TOM-Accuracy and TOM-RT indices. This index revealed a significant between-group difference ($t(26.3) = 2.23$, $p < .05$), which is also clear from visual inspection of the distributions (Fig. 6, left). To examine whether this composite could discriminate between autistic

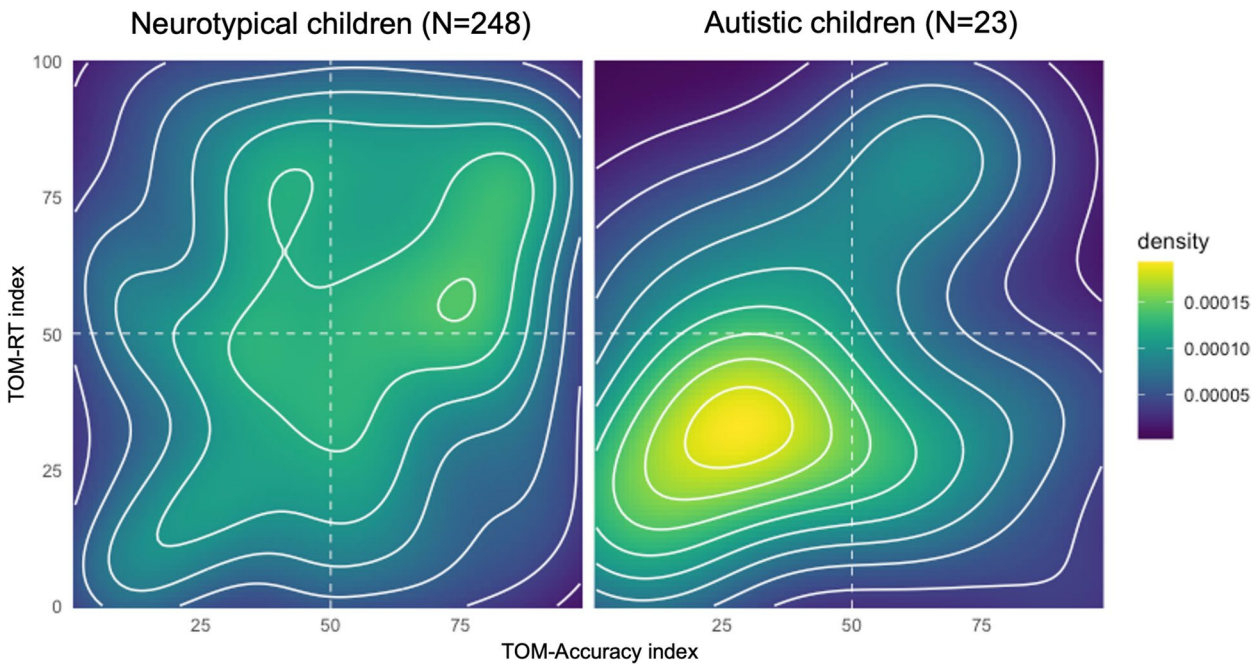

Fig. 5 Joint distribution of the TOM-Accuracy and the TOM-RT indices for the control group (left panel) and the autistic group (right panel)

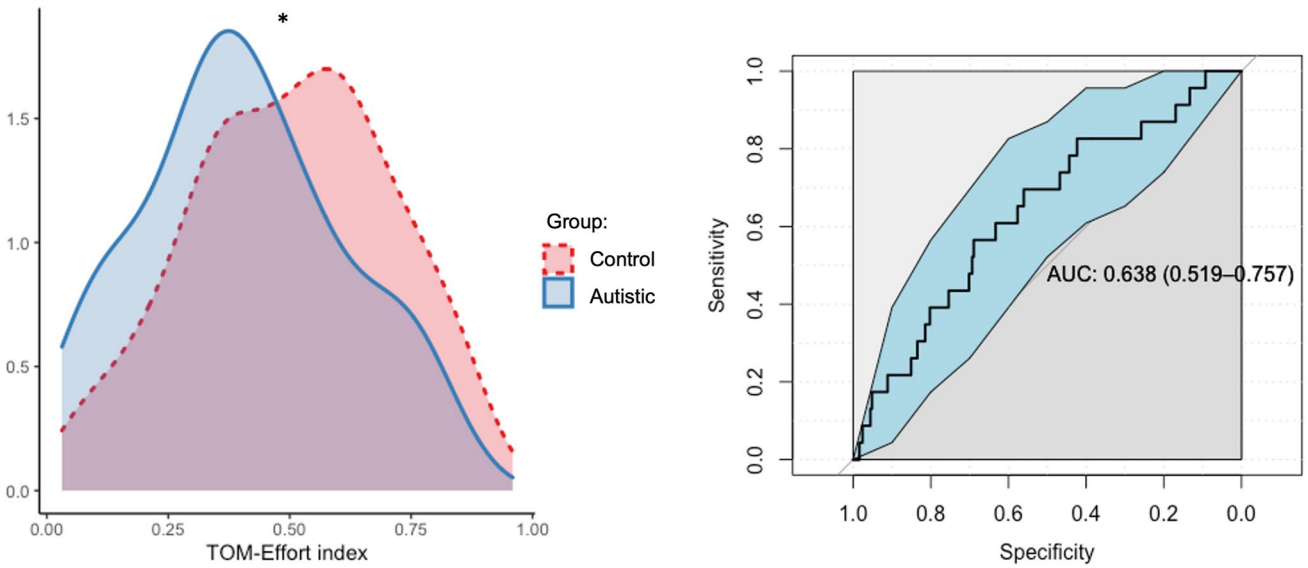

Fig. 6 Distribution across groups (left panel) and ROC curve (right panel) of the TOM-Effort index

and neurotypical children, we then performed a ROC analysis with this variable (Fig. 6, right). This revealed an area under curve (AUC) of .64 (95% *IC* = [.52-.76]), which was significantly different from chance but indicated only a moderate accuracy.

In the appendix, we report on three clinical cases that were extracted from our sample and which illustrate how this task can be used to provide rich clinical information at an individual level.

Discussion

In Experiment 2, the tablet-PST and the clinical indices built from Experiment 1 were applied to a group of autistic children. Before discussing the validity of the tablet-PST in

the General Discussion, we first briefly discuss the findings from this experiment at the group level in the context of the literature on autism and TOM. First, we note that the gender distribution is quite unbalanced in our autistic sample, which is common in autism [57]. However, Experiment 1 revealed that gender did not affect performance in this task, especially the TOM/non-TOM distinction. Therefore, such an imbalance should not obscure the interpretation of the data. An interesting aspect of our results was an interaction between item-type and group on accuracy, such that autistic children performed at higher rates than those in the neurotypical group on control items and at lower rates on false beliefs items. This dissociation is in line with what Baron-Cohen et al. [58] originally observed with another picture sequencing task in autistic children. It is also consistent with Binnie and Williams [59], who showed that when asked to sequence pictures, autistic children tended to prefer physical rather than psychological causalities, compared to neurotypical children. This dissociation led to the development of a more general empathizing/systemizing theory of autism [60] as a framework to explain the well-described TOM difficulties of autistic individuals. As for response times, we replicated Kaland et al. [26], who reported slowdowns on TOM items compared to controls, indicating a more pronounced effort in autism, yet only among younger participants in our case. While our sample size calls for caution regarding the interpretation of such subtle effects and for replication in future studies, it is arguably the case that the tablet-PST is not complex or sensitive enough to reveal such differences among older children. The pattern we observe also echoes the results from Zalla et al. ([61]) who observed (with a different picture sequencing task) that autistic children were typically slower in sequencing events involving characters' purposive action stakenbycharacters than physical events, while it was less the case among the neurotypical participants.

A novel aspect of our study concerns the use of clinical indices to evaluate TOM impairments at the individual level. Although the combined TOM-Effort index was the only one to show a significant difference between the autistic and the control groups, a visual inspection of the distribution of responses reveals a clear pattern in the autistic group that distinguishes itself from the pattern among neurotypical children. That is, TOM items were associated with lower accuracy and slower response times in the autistic group than among the neurotypical children. It is interesting that the combined TOM-Effort index revealed itself to be the most discriminating, since combining accuracy and response times has also been identified as a promising methods to measure TOM in autistic adults [25]. The ROC curve for that measure revealed that it could differentiate the autistic sample from the control group, although with moderate accuracy (AUC of 64%). This moderate level of accuracy was somewhat expected and confirms that TOM in general should not be considered *absent* in autism, but rather partial, delayed or simply atypical, the extent of that difference still being under debate (see [25], [20]). Such subtle between-group differences arguably explain why the combined index reached significance, whereas the accuracy and RT indices alone did not. Hence, discriminating autistic from non-autistic individuals should not be confounded as a validity measure of whether a given task assesses TOM [62].

Finally, Experiment 2 also provided evidence for adequate external validity by showing that the tablet-PST provided results that were consistent (in this group of children *without* language disorder) with a standard false belief task, in line with previous results with the PST in neurotypical children which had shown a correlation with social intelligence [37]; moreover, this association seemed to be specifically explained by TOM and not, say, by the general functioning of children, since no correlation appeared with a control measure of ADHD symptoms, as expected.

General Discussion

This study aimed to assess the suitability of a tablet adaptation of Langdon & Coltheart's [34] PST in order to ultimately assess children's TOM clinically and individually. This was achieved by testing a large sample of neurotypical children (Experiment 1), providing validation data and allowing for the construction of normalized clinical indices. As a test case, these indices were then applied to a sample of autistic children (Experiment 2).

Overall, our results provided evidence for the validity of the tablet-PST and in several different ways. Structurally, the CFA showed that the task behaved as it was intended. With respect to face and content validity, both the observations on the neurotypical and the autistic groups confirmed the validity of the task, since it behaved consistently with what can be expected from the TOM and the autism literature and, more specifically, with respect to the literature with paper-and-pencil-based presentations of the PST. Both experiments also supported the tablet-PST's discriminatory power, in the general population (with no floor or ceiling effects in the measures) and in a clinical context. We also showed, with respect to our autistic sample, that the tablet-PST provided consistent results with a standard false-belief task (external validity).

Unlike standard false-belief tasks, which typically rely on verbal material, the tablet-PST involves minimally verbal material and, as such, is more likely to be accessible to children with language difficulties. That said, we cannot rule out the possibility that participants might employ verbal strategies (e.g., inner speech) to complete the task. Therefore, verbal skills might still play a role in the performance with

this task (though the role of language might be reduced compared to other TOM tasks). Note, however, that this should affect both TOM and non-TOM items, and using the contrast between these conditions should control for that role, at least partly. Moreover, even when the interference of language with respect to *performance* in TOM tasks is ruled out, there is evidence supporting a relationship between language and TOM *competences* (see, e.g., [31]). In other words, the tablet-PST should probably not be expected to make TOM *independent* from language, but its minimally verbal material makes it suitable to investigate this complex relation in future research [63]. In addition to language, this task also controls for another possible confound observed in most TOM tasks, which is working memory (e.g. Arslan et al. [64]); indeed, with the tablet-PST all the necessary visual information remains displayed while the item is being completed. In terms of reliability, compared to the original PST, the inter-judge reliability of this tablet implementation is expected to be almost perfect, since the entire procedure and its scoring are automatized.

As far as applicability goes, the tablet format of the task showed many advantages with respect to data collection. Neurotypical children frequently reported that carrying out the task was enjoyable and their teachers described them as especially focused. This was also the case for autistic children, who remained engaged enough to complete the task, even when parents or clinicians had anticipated difficulties. These qualitative observations further support the clinical suitability of the task and confirm its usability in group settings [41].

The literature, as well as our results, shows that children should be expected to perform better on the control items compared to TOM items. However, our indices would allow clinicians to quantify *how much* better and to distinguish different profiles of participants, who might fail TOM items for different reasons. For example, children might fail TOM items because of a general difficulty in picture sequencing, which would be revealed by a low (control) GSA index; meanwhile, poor performance on TOM items could also result from normal sequencing abilities (attested by a good GSA index) combined with a specific TOM impairment (revealed by a poor TOM-Accuracy index). Similarly, by exploiting response times, our indices can distinguish between a child who performs in what is considered "typical" fashion for this task from another who performs correctly (in terms of accuracy) but exceptionally and effortfully, based on the TOM-RT index. Such reasoning (comparing control and test items, on the one hand, as well as accuracy and response times, on the other) is generally carried out by clinicians and is here directly computed by algorithms. As a consequence, the results of that reasoning present a reduced risk of human-based bias or between-clinician variability. Obviously, this is not to say that such calculations replace clinical reasoning, only that it provides data that is ultimately finer and more patient-calibrated than what raw performance scores could provide alone.

In terms of clinical utility, such data should help draw a more accurate cognitive and behavioral profile to inform interventions and assess their effects. When used during a diagnosis procedure however, this task should be used with much caution since, as we discussed in Experiment 2, the sensitivity analysis showed that autistic children should not be expected to systematically fail the tablet-PST nor should neurotypical children be expected to systematically succeed. For this reason, no specific clinical threshold was proposed.

We would be remiss if we did not underline three limitations as we consider future research. The first is that, while we had a large sample in Experiment 1, the sample size of our clinical group was limited. This is why Experiment 2 calls for replication and extension, including to other clinical groups, such as children with language disorders who usually show reduced TOM performance compared to neurotypical children, as shown in Nilsson & De López' [65] meta-analysis. However, as the authors of that study discuss, the extent to which the verbal nature of the tasks used contributes to this difference is still an open question. Future research should investigate this issue and would likely benefit from a task like the tablet-PST. Second, in Experiment 2 we compared the clinical sample to the neurotypical children who provided the normative data. Further validation would be provided by comparing this group to another independent group of neurotypical children. Third, the cross-sectional methodology we adopted did not allow for test-retest reliability assessment, which could be provided by a multiple testing or longitudinal paradigm.

In sum, after years of research on theory of mind using Langdon & Coltheart's [34] PST, our study investigates how a tablet-based implementation of this task could be naturally applied to individual clinical assessments. Our data, both in a large neurotypical sample and in a group of autistic children, supports the structural, face and content validity of the tablet-PST, as well as its external validity, sensitivity and clinical relevance. The results illustrate how this task can be helpful to assess the profile of children with atypical development, while providing normative data as background for it. This should benefit the assessment of autistic children but also other conditions that might encounter TOM difficulties, such as those with language disorders or psychiatric conditions.

Summary

Understanding others' minds (*theory of mind*, TOM) is an essential human trait. Disruptions of this ability are thus highly disabling and should be assessed correctly in clinical practice. Yet, most TOM tasks rely on verbal material, which

could conceivably amount to an obstacle for clinical populations as well as for clinicians who are interested in assessing them. One tool that avoids this limitation is Langdon and Coltheart's (1999) Picture Sequencing Task (PST), which is used in group-based research while using a minimally-verbal material. This work presents a tablet adaptation of this task with the aim of applying it to assess individuals in a clinical setting. Two of its advantages is that it is engaging to children, and it naturally allows for the collection of response times, which adds another dimension to analyses. We aimed to determine the extent to which this tablet-PST was a suitable assessment tool, and in two steps. In Experiment 1, we tested a large sample of neurotypical school-aged children (N = 248), in order to confirm that its results are comparable to what is found in paper-and-pencil based tasks in the literature. The results confirmed the task's structural and face validity. Based on these confirming group-level results, we then proposed three standardized clinical indices: i) a control index of the ability of a child to sequence pictures in general, ii) an accuracy index measuring how TOM can account for a child's accuracy, and iii) a response time index capturing the effort specifically associated with TOM items. In Experiment 2, these indices were applied to autistic children (N = 23), a clinical group that was expected to show atypical TOM performance. The data distributions showed that these children's outcomes were consistent with what is found in the autism literature and confirmed the task's clinical moderate sensitivity as well as its external validity. The tablet-PST thus appears as a suitable tool for assessing TOM in children generally, providing detailed cognitive profiles to inform clinical decisions.

Appendix

Clinical Vignettes

In order to illustrate how the tablet-PST can be used to provide clinical information, richer than a simple standardized score on a classical task, we extracted three prototypical profiles from our sample.

Participant 1 is a 6-years; 7-months-old boy. His autism diagnosis was confirmed by a positive SCQ and a positive ADOS-2. His level of intelligence functioning was also assessed and is strictly normal. In the tablet-PST, he obtained a GSA of 2.8/6, placing him only at the 1st percentile, given his age and SES. This score indicates that the task itself is not suitable to assess this specific child's skills. Indeed, the ability of that child to sequence pictures is compromised, maybe due to his young age, to behavioral difficulties and/or to difficulties in sequencing events and representing time which can be observed in certain individuals with autism [66]. If theory of mind must be assessed, it should be done otherwise.

Participant 7 presents a different profile. She is a 7-years; 6-months-old girl, whose autism diagnosis was confirmed by a positive SCQ and a positive ADI-R. Her level of intelligence functioning had not been formally tested, but no difficulties were suspected by the clinicians who follow her, and her language skills were in the normal to normal-superior range. In the tablet-PST, she got a 4.9/6 GSA (71st percentile) but scored 1.5/6 on average on TOM items (3rd percentile). Contrary to participant 1, participant 7 is thus very able to sequence pictures in general, with a GSA that is even in the normal-superior range. However, specific difficulties arise as soon as TOM is involved, where she scores much lower than expected, given her age and her good sequencing abilities. This is indicative of specific difficulties in theory of mind. No particularity was observed in her response times: she answered TOM items in 29 seconds on average, contra 23 seconds for the control items, which places her at percentile 26.

Finally, **participant 6** is a 7-years; 6-months-old boy, whose diagnosis was confirmed by a positive ADOS-2 and a positive SCQ. His intelligence was assessed as normal, as well as his language skills. He performed very well with the tablet-PST control items, with a GSA of 5.6/6 (85th percentile), and even better on TOM items with a mean accuracy of 6/6 (94th percentile). This child has no problem in sequencing pictures in general and, given his age and his good GSA, he performs even significantly better than expected with TOM items. However, he took 44 seconds on average to answer items involving false beliefs, contra only 18 seconds on control items: this difference, given his good accuracy, places him only at the 1st percentile. This indicated that participant 6's good accuracy on TOM items is obtained with unusual effort, in terms of response times, that is specific to items where TOM is involved. This might reflect atypical processing strategies such as compensatory mechanisms specifically associated with mentalizing.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10578-023-01648-0.

Acknowledgements We are to thank all the children who participated to this study, as well as Beynes school (Egletons, France) and Saint-Sacrement school (Lyon, France) which welcomed us despite a COVID context, Caroline Morellet and Isabelle Petit who helped with data collection in Experiment 1, and to Marie-Maude Geoffray as well as

to the teams of the Vinatier who helped with the recruitment of the clinical group of Experiment 2. This project was funded by the hospital Le Vinatier and the French Ministry of Health (PHRIP-2018-266).

Author Contributions NP performed conceptualization, funding acquisition, methodology, investigation, analysis, and wrote the original draft. IN performed supervision and writing—review and editing. MB performed writing—review and editing. JP provided resources and performed supervision, writing—review and editing.

Funding This work was supported by funds granted to the first author by the Centre Hospitalier Le Vinatier and the French Ministry of Health (PHRIP-2018-266).

Data Availability Data and analytic code to reproduce the analyses are available upon request to the corresponding author.

Declarations

Competing Interests The authors have no competing interests, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Ethical Approval All participants of these experiments provided informed consent to participate, as well as their parents. This study was run in accordance with the Declaration of Helsinki and received approval from the local IRB (Comité de Protection des Personnes Sud-Est I, ID RCB 2019-A01721-56).

References

1. Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1(4):515–526. https://doi.org/10.1017/S0140525X00076512
2. Waytz A, Gray K, Epley N, Wegner DM (2010) Causes and consequences of mind perception. Trends Cogn Sci 14(8):383–388. https://doi.org/10.1016/j.tics.2010.05.006
3. Frith U, Frith CD (2003) Development and neurophysiology of mentalizing. Philosophical Trans Royal Soc B: Biol Sci 358(1431):459–473. https://doi.org/10.1098/rstb.2002.1218
4. Bosco FM, Gabbatore I, Tirassa M (2014) A broad assessment of theory of mind in adolescence: the complexity of mindreading. Conscious Cogn 24:84–97. https://doi.org/10.1016/j.concog.2014.01.003
5. Dennett DC (1978) Beliefs about beliefs. Behav Brain Sci 1(4):568–570. https://doi.org/10.1017/S0140525X00076664
6. Wimmer H, Perner J (1983) Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13:103–128
7. Wellman HM, Cross D, Watson J (2001) Meta-analysis of theory-of-mind development: the truth about false belief. Child Dev 72(3):655–684. https://doi.org/10.1111/1467-8624.00304
8. Perner J, Wimmer H (1985) John thinks that Mary thinks that… attribution of second-order beliefs by 5- to 10-year-old children. J Exp Child Psychol 39(3):437–471. https://doi.org/10.1016/0022-0965(85)90051-7
9. O'Hare AE, Bremner L, Nash M, Happé F, Pettigrew LM (2009) A clinical Assessment Tool for Advanced theory of mind performance in 5 to 12 Year Olds. J Autism Dev Disord 39(6):916–928. https://doi.org/10.1007/s10803-009-0699-2
10. Beaudoin C, Leblanc É, Gagner C, Beauchamp MH (2020) Systematic review and inventory of theory of mind measures for Young Children. Front Psychol 10:2905. https://doi.org/10.3389/fpsyg.2019.02905
11. Hayward EO, Homer BD (2017) Reliability and validity of advanced theory-of-mind measures in middle childhood and adolescence. Br J Dev Psychol 35(3):454–462
12. Blair C, Razza RP (2007) Relating Effortful Control, executive function, and false belief understanding to emerging Math and literacy ability in Kindergarten. Child Dev 78(2):647–663. https://doi.org/10.1111/j.1467-8624.2007.01019.x
13. Dunn J, Cutting AL, Fisher N (2002) Old friends, New friends: predictors of children's perspective on their friends at School. Child Dev 73(2):621–635. https://doi.org/10.1111/1467-8624.00427
14. Fink E, Begeer S, Hunt C, de Rosnay M (2014) False-belief understanding and social preference over the first 2 years of school: a longitudinal study. Child Dev 85:2389. https://doi.org/10.1111/cdev.12302
15. Maat A, Fett A-K, Derks E (2012) Social cognition and quality of life in schizophrenia. Schizophr Res 137(1):212–218. https://doi.org/10.1016/j.schres.2012.02.017
16. Bivona U, Formisano R, Laurentiis S, Accetta N, Cosimo M, Massicci R, Ciurli P, Azicnuda E, Silvestro D, Sabatini U, Falletta Caravasso C, Carlesimo G, Caltagirone C, Costa A (2015) Theory of mind impairment after severe traumatic brain injury and its relationship with caregivers' quality of life. Restor Neurol Neurosci 33:335. https://doi.org/10.3233/RNN-140484
17. Botting N, Conti-Ramsden G (2008) The role of language, social cognition, and social skill in the functional social outcomes of young adolescents with and without a history of SLI. Br J Dev Psychol 26(2):281–300. https://doi.org/10.1348/026151007X235891
18. Baron-Cohen S (1997) Mindblindness: an essay on autism and theory of mind. MIT press, Cambridge
19. Fletcher-Watson S, Happé F (2019) Autism: a new introduction to psychological theory and current debate. Routledge, London
20. Marocchini E (2023) Impairment or difference? The case of theory of mind abilities and pragmatic competence in the Autism Spectrum. Appl Psycholinguist 44:1–19. https://doi.org/10.1017/S0142716423000024
21. Senju A (2012) Spontaneous theory of mind and its absence in Autism Spectrum disorders. The Neuroscientist 18(2):108–113. https://doi.org/10.1177/1073858410397208
22. Baltazar M, Geoffray M, Chatham C, Bouvard M, Martinez Teruel A, Monnet D, Scheid I, Murzi E, Couffin-Cadiergues S, Umbricht D, Murtagh L, Delorme R, Ly Le-Moal M, Leboyer M, Amestoy A (2021) Reading the mind in the eyes in autistic adults is modulated by Valence and Difficulty: an InFoR Study. Autism Res 14(2):380–388. https://doi.org/10.1002/aur.2390
23. Livingston LA, Happé F (2017) Conceptualising compensation in neurodevelopmental disorders: reflections from autism spectrum disorder. Neurosci Biobehavioral Reviews 80:729–742. https://doi.org/10.1016/j.neubiorev.2017.06.005
24. Behrmann M, Avidan G, Leonard GL, Kimchi R, Luna B, Humphreys K, Minshew N (2006) Configural processing in autism and its relationship to face processing. Neuropsychologia 44(1):110–129. https://doi.org/10.1016/j.neuropsychologia.2005.04.002
25. Livingston LA, Carr B, Shah P (2019) Recent advances and new directions in measuring theory of mind in autistic adults. J Autism Dev Disord 49(4):1738–1744. https://doi.org/10.1007/s10803-018-3823-3
26. Kaland N, Smith L, Mortensen EL (2007) Response Times of Children and adolescents with Asperger Syndrome on an 'Advanced' test of theory of mind. J Autism Dev Disord 37(2):197–209. https://doi.org/10.1007/s10803-006-0152-8

27. Happé F (1994) An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. J Autism Dev Disord 24(2):129–154. https://doi.org/10.1007/BF02172093
28. Baron-Cohen S, O'Riordan M, Stone V, Jones R, Plaisted K (1999) Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. J Autism Dev Disord 29:407
29. Baron-Cohen S, Wheelwright S, Hill J, Raste Y, Plumb I (2001) The reading the mind in the eyes test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. J Child Psychol Psychiat 42(2):241–251
30. de Villiers J (2007) The interface of language and theory of mind. Lingua 117(11):1858–1878. https://doi.org/10.1016/j.lingua.2006.11.006
31. Milligan K, Astington JW, Dack LA (2007) Language and Theory of mind: Meta-Analysis of the Relation between Language ability and false-belief understanding. Child Dev 78(2):622–646. https://doi.org/10.1111/j.1467-8624.2007.01018.x
32. Pyers JE, Senghas A (2009) Language promotes false-belief understanding: evidence from learners of a New sign Language. Psychol Sci 20(7):805–812. https://doi.org/10.1111/j.1467-9280.2009.02377.x
33. Levy SE, Giarelli E, Lee L-C, Schieve LA, Kirby RS, Cunniff C, Nicholas J, Reaven J, Rice CE (2010) Autism spectrum disorder and co-occurring Developmental, Psychiatric, and medical conditions among children in multiple populations of the United States. J Dev Behav Pediatr 31(4):267. https://doi.org/10.1097/DBP.0b013e3181d5d03b
34. Langdon R, Coltheart M (1999) Mentalising, schizotypy, and schizophrenia. Cognition 71(1):43–71. https://doi.org/10.1016/S0010-0277(99)00018-9
35. Payne JM, Porter M, Pride NA, North KN (2016) Theory of mind in children with neurofibromatosis type 1. Neuropsychology 30(4):439–448. https://doi.org/10.1037/neu0000262
36. Porter MA, Coltheart M, Langdon R (2008) Theory of mind in Williams Syndrome assessed using a Nonverbal Task. J Autism Dev Disord 38(5):806–814. https://doi.org/10.1007/s10803-007-0447-4
37. Rajkumar AP, Yovan S, Raveendran AL, Russell PSS (2008) Can only intelligent children do mind reading: the relationship between intelligence and theory of mind in 8 to 11 years old. Behav Brain Funct 4(1):51. https://doi.org/10.1186/1744-9081-4-51
38. Van Rheenen TE, Rossell SL (2013) Picture sequencing task performance indicates theory of mind deficit in bipolar disorder. J Affect Disord 151(3):1132–1134. https://doi.org/10.1016/j.jad.2013.07.009
39. Piatt C, Coret M, Choi M, Volden J, Bisanz J (2016) Comparing children's performance on and preference for a number-line estimation Task: Tablet Versus Paper and Pencil. J Psychoeducational Assess 34(3):244–255. https://doi.org/10.1177/0734282915594746
40. Lane R, Radesky J (2019) Digital Media and Autism Spectrum disorders: review of evidence, theoretical concerns, and opportunities for intervention. J Dev Behav Pediatrics: JDBP 40(5):364–368. https://doi.org/10.1097/DBP.0000000000000664
41. Bignardi G, Dalmaijer ES, Anwyl-Irvine A, Astle DE (2021) Collecting big data with small screens: group tests of children's cognition with touchscreen tablets are reliable and valid. Behav Res Methods 53(4):1515–1529. https://doi.org/10.3758/s13428-020-01503-3
42. Westby C, Robinson L (2014) A developmental perspective for promoting theory of mind. Top Lang Disorders 34(4):362–382
43. Van Hof 'T, Tisseur M, Van Berckelear-Onnes C, Van Nieuwenhuyzen I, Daniels A, Deen AM, Hoek M, H. W., Ester WA (2021) Age at autism spectrum disorder diagnosis: a systematic review and meta-analysis from 2012 to 2019. Autism 25(4):862–873. https://doi.org/10.1177/1362361320971107
44. Currie C, Molcho M, Boyce W, Holstein B, Torsheim T, Richter M (2008) Researching health inequalities in adolescents: the development of the Health Behaviour in School-aged children (HBSC) family affluence scale. Soc Sci Med 66(6):1429–1436. https://doi.org/10.1016/j.socscimed.2007.11.024
45. R Core Team (2022) R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/
46. Rosseel Y (2012) lavaan: An R package for structural equation modeling. J Stat Softw 48(2):1–36. https://doi.org/10.18637/jss.v048.i02
47. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear mixed-effects models using lme4. J Stat Softw 67(1):1–48. https://doi.org/10.18637/jss.v067.i01
48. Kuznetsova A, Brockhoff PB, Christensen RH (2017) lmerTest package: tests in linear mixed effects models. J Stat Softw 82:1–26
49. Lenth R (2022) emmeans: Estimated Marginal Means, aka Least-Squares Means ({R package version 1.8.2}) [Computer software]. https://CRAN.R-project.org/package=emmeans
50. Crawford JR, Garthwaite PH (2006) Comparing patients' predicted test scores from a regression equation with their obtained scores: a significance test and point estimate of abnormality with accompanying confidence limits. Neuropsychology 20(3):259–271. https://doi.org/10.1037/0894-4105.20.3.259
51. Wellman HM (2018) Theory of mind: the state of the art. Eur J Dev Psychol 15(6):728–755. https://doi.org/10.1080/17405629.2018.1435413
52. Baron-Cohen S (2002) The extreme male brain theory of autism. Trends Cogn Sci 6(6):248–254
53. Charman T, Ruffman T, Clements W (2002) Is there a gender difference. False Belief Development? Social Development 11(1):1–10. https://doi.org/10.1111/1467-9507.00183
54. Launay L, Maeder C, Roustit J, Touzin M et al (2018) Évaluation du Langage écrit et du Langage Oral 6-15 ans (EVALEO 6-15). OrthoÉdition 40:43–51
55. Mercier C, Roche S, Gaillard S, Kassai B, Arzimanoglou A, Herbillon V, Roy P, Rheims S (2016) Partial validation of a French version of the ADHD-rating scale IV on a French population of children with ADHD and Epilepsy. Factorial structure, reliability, and responsiveness. Epilepsy Behav 58:1–6. https://doi.org/10.1016/j.yebeh.2016.02.016
56. Montemurro S, Daini R, Tagliabue C, Guzzetti S, Gualco G, Mondini S, Arcara G (2022) Cognitive reserve estimated with a life experience questionnaire outperforms education in predicting performance on MoCA: Italian normative data. Curr Psychol. https://doi.org/10.1007/s12144-022-03062-6
57. Fombonne E (2009) Epidemiology of pervasive developmental disorders. Pediatr Res 65(6):591–598
58. Baron-Cohen S, Leslie AM, Frith U (1986) Mechanical, behavioural and intentional understanding of picture stories in autistic children. Br J Dev Psychol 4(2):113–125. https://doi.org/10.1111/j.2044-835X.1986.tb01003.x
59. Binnie L, Williams J (2003) Intuitive psychology and physics among children with autism and typically developing children. Autism 7(2):173–193. https://doi.org/10.1177/1362361303007002005
60. Baron-Cohen S (2009) Autism: the empathizing–systemizing (E-S) theory. Annals of the New York. Acad Sci 1156(1):68–80. https://doi.org/10.1111/j.1749-6632.2009.04467.x
61. Zalla T, Labruyere N, Georgieff N (2006) Goal-Directed Action representation in Autism. J Autism Dev Disord 36(4):527–540. https://doi.org/10.1007/s10803-006-0092-3

62. Chevallier C (2012) Theory of mind and autism: revisiting Baron-Cohen et al.'s Sally-Anne study. Developmental psychology: revisiting the classic studies. Sage Publications Ltd, Thousand Oaks, pp 148–163
63. Bosco FM, Tirassa M, Gabbatore I (2018) Why pragmatics and theory of mind do not (completely) overlap. Front Psychol. 9https://doi.org/10.3389/fpsyg.2018.01453
64. Arslan B, Hohenberger A, Verbrugge R (2017) Syntactic recursion facilitates and Working Memory predicts recursive theory of mind. PLoS ONE 12(1):e0169510. https://doi.org/10.1371/journal.pone.0169510
65. Nilsson KK, De López KJ (2016) Theory of mind in Children with Specific Language Impairment: a systematic review and Meta-analysis. Child Dev 87(1):143–153. https://doi.org/10.1111/cdev.12462
66. Jurek L, Longuet Y, Baltazar M, Amestoy A, Schmitt V, Desmurget M, Geoffray M-M (2019) How did I get so late so soon? A review of time processing and management in autism. Behav Brain Res 374:112121. https://doi.org/10.1016/j.bbr.2019.112121
67. Botha M, Hanlon J, Williams GL (2021) Does Language Matter? Identity-first Versus Person-First Language Use in Autism Research: a response to Vivanti. J Autism Dev Disord. https://doi.org/10.1007/s10803-020-04858-w
68. Bury SM, Jellett R, Spoor JR, Hedley D (2020) "It defines who I am" or "It's something I have": What language do [autistic] Australian adults [on the autism spectrum] prefer? J Autism Dev Disord 53:677–687