Topics in Cognitive Science 0 (2022) 1–30
© 2022 Cognitive Science Society LLC
ISSN: 1756-8765 online
DOI: 10.1111/tops.12597

This article is part of the topic "Individual Differences in Spatial Navigation: Building a Cognitive Science of Human Variation," Nora S. Newcombe, Mary Hegarty, and David Uttal (Topic Editors).

Measuring Spatial Perspective Taking: Analysis of Four Measures Using Item Response Theory

Maria Brucato,[a] Andrea Frick,[b] Stefan Pichelmann,[b] Alina Nazareth,[c] Nora S. Newcombe[a]

[a]*Department of Psychology, Temple University*
[b]*Department of Psychology, University of Fribourg*
[c]*Meta*

Received 3 March 2021; received in revised form 3 December 2021; accepted 6 December 2021

Abstract

Research on spatial thinking requires reliable and valid measures of individual differences in various component skills. Spatial perspective taking (PT)—the ability to represent viewpoints different from one's own—is one kind of spatial skill that is especially relevant to navigation. This study had two goals. First, the psychometric properties of four PT tests were examined: Four Mountains Task (FMT), Spatial Orientation Task (SOT), Perspective-Taking Task for Adults (PTT-A), and Photographic Perspective-Taking Task (PPTT). Using item response theory (IRT), item difficulty, discriminability, and efficiency of item information functions were evaluated. Second, the relation of PT scores to general intelligence, working memory, and mental rotation (MR) was assessed. All tasks showed good construct validity except for FMT. PPTT tapped a wide range of PT ability, with maximum measurement precision at average ability. PTT-A captured a lower range of ability. Although SOT contributed less measurement information than other tasks, it did well across a wide range of PT ability. After controlling for general intelligence and working memory, original and IRT-refined versions of PT tasks were each related to MR. PTT-A and PPTT showed relatively more divergent validity from MR than SOT. Tests of dimensionality indicated that PT tasks share one common PT dimension, with secondary task-specific factors also impacting the measurement of individual differences in performance.

Correspondence should be sent to Maria Brucato, Department of Psychology, Temple University, Philadelphia, PA 19122, USA. E-mail: mbrucato@temple.edu

Advantages and disadvantages of a hybrid PT test that includes a combination of items across tasks are discussed.

Keywords: Spatial perspective taking; Spatial cognition; Individual differences; Item response theory

1. Introduction

Measuring spatial perspective taking: Analysis of four measures using item response theory (IRT) Cognitive scientists aim to describe the working of the human mind and brain, typically with the implicit or even explicit assumption that there is fundamental commonality across individuals and cultures. In this view, variation among people in behavioral or neural data is annoying error variance, not vital data to use in constraining models and theories. Although Cronbach (1957) highlighted the tension between the normative approach and the study of individual differences long ago, in his classic article on "The Two Disciplines of Scientific Psychology," the field has done little to close the gap in the decades following.

Addressing individual differences requires cognitive scientists to change their research practices in several ways. First, we need to address how to design measures and paradigms that have adequate psychometric characteristics. Social and personality psychologists have been facing up this challenge for some time and continue to make progress (Flake, Pek, & Hehman, 2017; Hussey & Hughes, 2020). Cognitive science is only beginning to take up the challenge (Draheim, Tsukahara, Martin, Mashburn, & Engle, 2020). There is growing recognition that many experimental cognitive tasks are designed to minimize variability (such as the Stroop test), and many others have unevaluated psychometric characteristics. That is, they may not be internally consistent or offer acceptable test–retest reliability. Second, we need to recognize the theoretical purchase we gain by studying how people differ, as well as age-related change and cultural variation. Strikingly, researchers in language and language acquisition, an area that has usually deemphasized individual differences, are beginning to adopt this strategy (Kidd, Donnelly, & Christiansen, 2018).

Spatial cognition is one example of an area of psychology where fundamental concepts do not yet have reliable and valid measures. There are many kinds of spatial skills (Newcombe & Shipley, 2015; Uttal & Cohen, 2012). Back in 1996, Lohman emphasized the importance of constructing tests that measure distinct aspects of spatial ability by emphasizing component skills. One key skill is spatial perspective taking (PT), or the ability to represent a viewpoint different from one's own (i.e., "spatial orientation"; Lohman, 1979). A basic form of PT, called Level 1 and established early in life, entails simply understanding that other people have different viewpoints and perceive objects along their own lines of sight. In Level 2 PT, people precisely imagine these different viewpoints, which is much more challenging (Flavell, Flavell, Green, & Wilcox, 1981; Lempers, Flavell, & Flavell, 1977; Masangkay et al., 1974).

PT is historically underinvestigated in the spatial-reasoning literature, with most research focused on skills pertaining to object-based transformations (e.g., mental rotation; MR) as opposed to reference-frame transformations (i.e., PT). However, recent work has demonstrated the additional predictive power of PT in addition to MR for navigation and cognitive map building (Fields & Shelton, 2006; Kozhevnikov, Motes, Rasch, & Blajenkova, 2006; Nazareth, Weisberg, Margulis, & Newcombe, 2018). In addition, PT may contribute to success in science, technology, engineering, and mathematics (STEM) disciplines that benefit from the ability to adopt a survey view of the world, such as astronomy, geoscience, or architecture (e.g., Nazareth, Newcombe, Shipley, Velazquez, & Weisberg, 2019; Oldakowski, 2001; Plummer, Bower, & Liben, 2016; Wai, Lubinski, & Benbow, 2009).

Although PT and MR are computationally similar in that they are both dynamic spatial skills (support movement/transformations in space), MR is thought to rely on object-based transformations whereas PT relies on perspective-based transformations (i.e., maintaining spatial relationships between allocentric cues in the environment while shifting one’s frame of reference; Hegarty & Waller, 2004; Lohman, 1979). There is some evidence that distinct neural substrates support these two processes: brain activity in the parieto-occipital sulcus/retrosplenial cortex and the hippocampus increases when participants are instructed to perform a perspective-based spatial transformation as opposed to an object-based transformation (Committeri et al., 2004; Lambrey, Doeller, Berthoz, & Burgess, 2012). In behavioral work, there is evidence that MR and PT vary independently across individuals and factor-analytic models fit behavioral data best via separate as opposed to combined factors (Amorim & Stucchi, 1997; Frick, 2019; Hegarty & Waller, 2004; Huttenlocher & Presson, 1973; Kozhevnikov & Hegarty, 2001; Pittalis & Christou, 2010; Wraga, Creem, & Proffitt, 2000; Zacks, Mires, Tversky, & Hazeltine, 2000) although the skills are often very highly correlated.

Over the past decade, investigators have developed several PT assessments, many of which took inspiration from the “three mountains task”—one of the earliest measures of PT (Piaget & Inhelder, 1956). However, existing assessment tools vary substantially in stimuli and paradigms. This can be beneficial for the generalization of study results (Schmiedek, Lövdén, & Lindenberger, 2014), but can also cause confusion. Specifically, tests vary in complexity (detail and number of allocentric cues within the environment), agency (presence of an agent or not), and memory load (memory of a target scene/image required or not).

Each of these factors may substantially affect the nature of the ability assessed. First, the number of allocentric cues within the stimulus environment may be important because PT tasks with more cues offer the possibility of using external landmarks rather than spatially updating an egocentric reference frame (Burgess, Spiers, & Paleologou, 2004; Frick, Möhring, & Newcombe, 2014; Lambrey et al., 2012). Whether allocentric cues are two-dimensional or three-dimensional impacts the speed at which individuals with various levels of spatial ability can solve transformation problems (Cooper & Regan, 1982 as cited in Lohman, 1996). Second, the presence of agents (also called avatars) in stimuli may be a source of unintentional task variance related to social cognition. A recent set of studies found that when taking the spatial perspective of avatars such as humans or even dolls (but not objects perceived to have no agency), individuals with weaker social skills were significantly less accurate than those with strong social skills (Clements-Stephens, Vasiljevic, Murray, & Shelton, 2013; Shelton, Clements-Stephens, Lam, Pak, & Murray, 2012). Third, PT in the context of survey-based navigation often requires visualization of perspectives recalled from memory (e.g., when one plans a return route and recalls from memory what it looked like on the way there, and then adjusts the perspective 180°). Thus, memory load may be an important factor for an ecologically valid PT measure. On the other hand, PT tasks that tax memory make it more difficult to disentangle whether individual differences in performance are due to PT ability or memory ability. Therefore, a PT task that does not tax memory may be a more process-pure measure of the spatial skill.

The present study examined several psychometric properties of four PT tasks. Given the vast range of PT tasks and paradigms, we selected four PT tasks to assess in the present study based on several criteria. First, we only included measures that were formatted as tests of Level 2 PT rather than experimental paradigms (e.g., Aichhorn, Perner, Kronbichler, Staffen, & Ladurner, 2006; Burles, Slone, & Iaria, 2017; Vogeley et al., 2004). Second, we only included measures that had more than one object in the scene to encourage PT rather than MR. Finally, we sought out tasks that were representative of each category of a 2 × 2 design with one factor being agency (i.e., whether the task asked participants to imagine a scene from their own point-of-view or that of an agent) and the second factor being the type of allocentric cues (i.e., whether the task stimuli depicted two-dimensional or three-dimensional objects; cf., Fig. 2).

First, we investigated if a convergent underlying ability is measured within and between all four tasks by testing assumptions of unidimensionality and comparing fit statistics of competing models of dimensional structure, using IRT. We also evaluated the level of measurement precision for each task along a range of PT ability and explored options for optimizing task efficiency in the presence of redundant or minimally informative items. Second, we assessed the relation between PT scores and general intelligence, working memory, and MR, to establish discriminant validity of the PT measures with respect to MR as a related but dissociable ability. Third, we tested the reliability of Photographic Perspective Taking Task (PPTT; Plummer et al., 2016) and Perspective Taking Task for Adults (Frick et al., 2014) originally used with children that were adapted in difficulty for use with adults. Reliabilities for the two other tasks, Spatial Orientation Task ($\alpha = .88$; Kozhevnikov & Hegarty, 2001) and Four Mountains Task (intraclass coefficient = .81; Chan et al., 2016) have been previously reported as high. Finally, we assessed the need for and feasibility of a hybrid PT test containing a combination of best items across PT tasks in the event that (a) PT tasks tended to measure different kinds of abilities, (b) a hybrid task offered greater divergent validity from MR, or (c) a hybrid task provided greater measurement information for a wider range of PT ability than individual tasks alone.

2. Advantages of item response theory

IRT is a modern psychometric paradigm, which models the probability of observing a response to an item in a test, questionnaire, or task. Traditionally stemming from the education literature to optimize test scoring and computerized adaptive testing, the approach has gained traction in psychology for psychometric examination of measures of complex cognitive constructs. IRT holds several advantages over traditional psychometric paradigms like classical test theory in that (a) item parameter estimates are sample-independent and more stable across task versions calibrated on the same scale, (b) the estimation of a task's measurement precision takes into account the ability level of an individual rather than assuming it to be equal across individuals, and (c) the potential for guessing as opposed to an ability-informed response can be considered (Fan, 1998; Jabrayilov, Emons, & Sijtsma, 2016; Magno, 2009). Thus, the use of IRT allows for greater confidence that item parameters are generalizable to the larger population beyond our current sample, are stable if a combined battery of items between tasks is formed, and are more precise by considering the interaction of item difficulty with high- or low-level ability.

Several types of IRT models for estimating the association between an individual's dichotomous response to an item (e.g., correct or incorrect) and their ability level exist (Thissen & Steinberg, 1986). Rasch (1960) proposed a logistic function of the difference between the ability of an individual and the difficulty of an item so that "specific objectivity" could be achieved. That is, the independence of item parameter estimates from the abilities of a specific sample of individuals is maximized and vice versa. Because of these features, logistic IRT models are used in the present study. Specifically, we opt to use two-parameter logistic (2PL) models because of these estimate parameters for both discriminability and difficulty of items (Lord & Novick, 2008).

Another advantage of the IRT approach is that the dimensionality of a set of items can be assessed by comparing fit parameters of competing IRT models of varying dimensions. Thus, in line with our aims above, we tested three competing models of dimensionality in the current study. Model A (see Fig. 1) is a unidimensional model wherein PT items from all tasks load onto one general PT ability. This model was used to test a structure in which a general PT dimension was the only factor that affected responses on items from all tasks. Model B is a non-hierarchical between-item multidimensional IRT model (MIRT; Adams, Wilson, & Wang, 1997; Liu, Magnus, O'Connor, & Thissen, 2018), in which there are two or more primary abilities that are correlated with each other, and any given item in a measure is allowed to load positively onto only one ability. This MIRT model was used to test a structure in which each PT task measured a distinct ability which impacted responses on only items within it, and these distinct abilities were correlated (Fig. 1b). Alternatively, Model C is a hierarchical bifactor MIRT model (Gibbons & Hedeker, 1992) that allows all items to load positively onto one factor (representing one *general* ability) and additionally onto zero, one, or more alternate factors, which are not correlated with each other nor with the primary ability. This MIRT model was used to test a structure in which a primary PT dimension affected responses on items from all tasks, while other dimensions unrelated to PT that varied between tasks additionally contributed to responses on all items of only some tasks. A comparison of fit among competing Models A, B, and C was used to inform a robust and valid measure of PT, similar to previous subscale formation for questionnaires assessing complex cognitive processes (Gibbons, Rush, & Immekus, 2009; Reise, Morizot, & Hays, 2007).

For testing these models, we formulated the following predictions: We expected that, since items within each task did not differ much with respect to stimuli and paradigm, each task would meet the unidimensionality and local dependence assumptions and show a good fit to a unidimensional IRT model. Conversely, given that there were clear differences in stimuli and paradigms between the four PT tasks, we predicted that the assumptions of local dependence

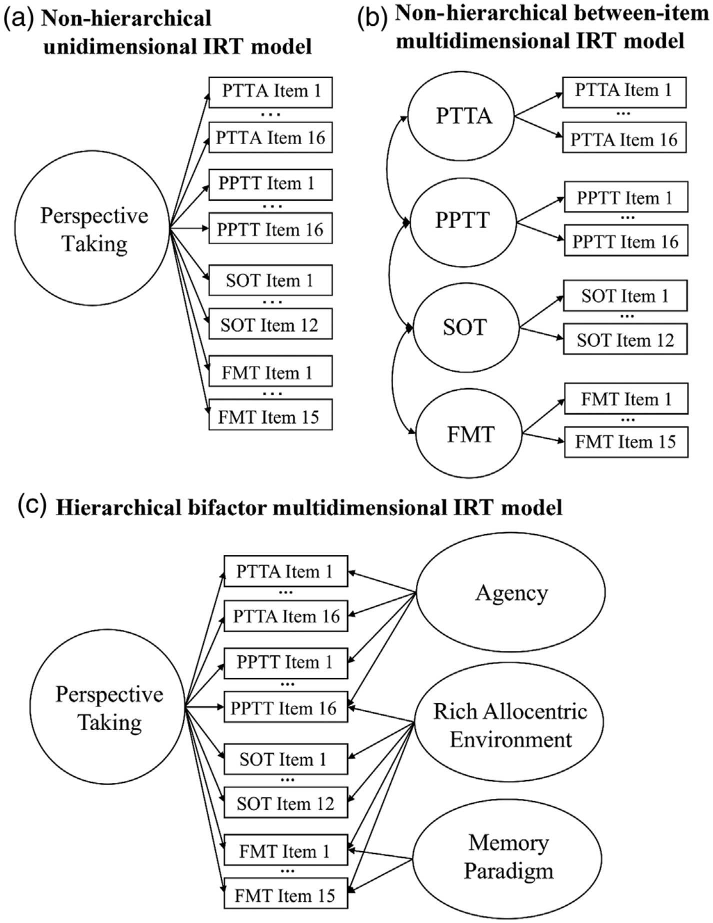

Fig 1. Competing models for perspective taking dimensionality. Abbreviations: PPTT, Photographic Perspective Taking Task; PTTA, Perspective Taking Task for Adults; SOT, Spatial Orientation Task; FMT, Four Mountains Task.

and unidimensionality would be violated when attempting to fit a unidimensional model to a combined battery of items of all tasks. In this case, violation of unidimensionality can be an indicator of either superfluous task features or meaningful dimensions of PT. Therefore, we tested three possibilities for dimensional structure: (1) if all tasks measure a common uniform PT ability which is uninfluenced by task-dependent features, then Model A should show the best fit to the data, (2) if all tasks measure a common PT ability, but there are additional uncorrelated dimensions which contribute to individual differences in responses, then Model C should show the best fit, or (3) if each task measures distinct but related abilities, then Model B should show the best fit. Finally, we predicted that performance on a refined composite PT measure that was created on the basis of the IRT analysis would provide a more accurate estimate of PT, and therefore allow us to distinguish it more clearly from MR when controlling for general intelligence and working memory than an unrefined composite PT battery of all items or any single task alone.

3. Methods

3.1. Participants

Session 1 data were available for 110 Temple University undergraduates (86 female; *M* age = 20.21 years, *SD* = 3.28) who volunteered to participate for course credit. The initial sample included 135 people, but data from 25 participants were discarded (12 participants did not report English fluency before age five, two participants showed inattentive responding as indicated by failure of all in-task comprehension checks, two participants had insufficient data due to experimenter error, and nine had insufficient data due to computer error). All participants had a normal or corrected-to-normal vision. Participants in session 1 were invited to return for session 2 for $15 compensation, and 48 participants did so (36 female; *M* age = 19.96, *SD* age = 1.87).

3.2. Procedure

The study consisted of two 1-h sessions separated by at least one but no more than 2 weeks. Session 2 data were used only to conduct test–retest reliability analysis and to assess drop-out effects. Informed consent was obtained at each session. Table 1 lists the fixed order of tasks at each session. All measures were administered on two Windows 7 64-bit computers with 40 × 62-cm LCD monitor display, except for the general intelligence measures, which were administered via a testing booklet. Up to two participants were tested at once in our testing space, with one experimenter per participant. Participants were randomly assigned to one of two testing computers for each session. All computer tasks were silent, and computers were separated by cubicles on opposite sides of the room, so participants could not see nor hear each other and there were no distractions. After completion of computer tasks, experimenters brought their participants into a separate private testing room where the verbal fluency or IQ task was administered.

3.3. Measures of perspective taking

3.3.1. Photographic Perspective Taking Task

The PPTT(Plummer et al., 2016) was adapted for use in adults by limiting time for a response on each item to 10 s. The task was also adapted from a paper-and-pencil task to a computerized version. Participants saw an image of a three-dimensional doll viewing a display of two differently colored two-dimensional circles (see Fig. 2a). Eight different displays were simultaneously provided below this image showing the circles from different viewpoints in 45° increments. Participants were asked to determine which one showed what someone would see from where the doll was standing, and there was no direct instruction for strategy

Table 1
Descriptive statistics for all measures and task order

Construct	Measure	Minimum	Maximum	*M*	*SD*	*Skewness*
	Session One					
PT	Photographic Perspective Taking Task	0.13	1	0.69	0.23	−0.47
-	Demographics Questionnaire					
PT	Perspective Taking Task for Adults	0.06	0.94	0.54	0.19	−0.32
MR	Mental Rotation Test (Peters)	0	23	8.95	4.93	0.60
PT	Four Mountains Task	0.27	1	0.71	0.15	−0.31
SWM	Symmetry Span	6	41	27.57	7.54	−0.56
PT	Spatial Orientation Task	0	1	0.55	0.25	−0.24
GI	Wide Range Achievement Test – Reading	0.42	0.98	0.84	0.09	−1.38
	Session Two					
PT	Photographic Perspective Taking Task	0.06	1	0.77	0.21	−1.07
SWM	Position Span Task	1	27	13.5	5.83	−0.67
PT	Perspective-Taking Task for Adults	0.13	0.9	0.6	0.2	−0.66
MR	Mental Rotation Test (Ganis and Kievit)	17	46	28.6	5.56	1.14
GI	WASI-II Full-Scale 2-Subtest IQ	79	117	98.23	8.82	−0.06

Note. Measures are listed in the order that they were given during the experiment.

Abbreviations: PT, perspective taking; MR, mental rotation; SWM, spatial working memory; GI., general intelligence; WASI-II, Weschler Abbreviated Scale of Intelligence.

use. There was one practice item followed by 13 test items presented in random order for each participant at each session. Test items varied in the arrangement of the colored circles and the vantage point of the doll around the display in 45° increments. Three additional items in the test, where the doll and participant's perspective were the same, were comprehension checks for attentive responding such that incorrect responses to all three items resulted in exclusion from the analysis. Aside from this, first-person perspective (1PP) items were not used in any of the following analyses. The 90% completion rate (proportion of participants that responded to at least 12 out of 13 test items) was 0.87.

3.3.2. *Perspective-Taking Task for Adults*

A computerized version of the Perspective-Taking Task for Children (PTT-C; Frick et al., 2014) was adapted for use in adults, by using only layouts with three objects (as opposed to fewer) and by presenting a larger number of more difficult response alternatives. Participants saw an image of a three-dimensional figurine taking a photograph of an arrangement of three different-colored three-dimensional objects (see Fig. 2b). Eight different object-arrangements were simultaneously provided below this image, and participants were asked to determine which one looked like the picture the figurine could have taken from where it was standing.

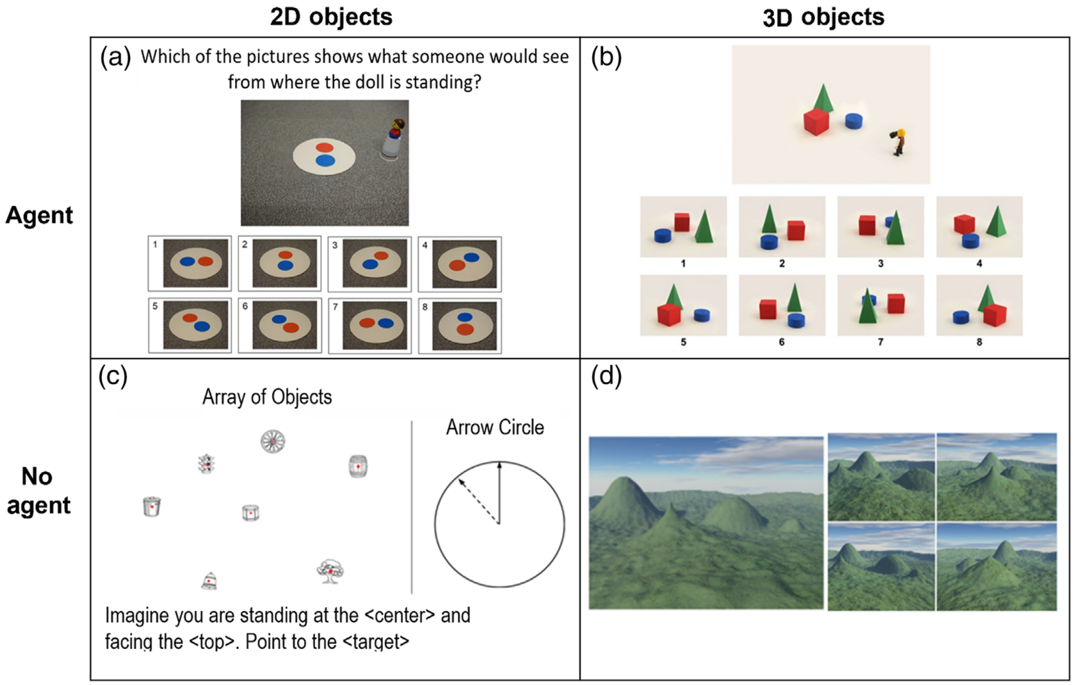
2D objects
3D objects
Agent
No agent
(a) Which of the pictures shows what someone would see from where the doll is standing?
1
2
3
4
5
6
7
8
(b)
1
2
3
4
5
6
7
8
(c)
Array of Objects
Arrow Circle
Imagine you are standing at the <center> and facing the <top>. Point to the <target>
(d)

Fig 2. Four measures of perspective taking: (a) Photographic Perspective-Taking Task. (b) Perspective-Taking Task for Adults. (c) Spatial Orientation Task. (d) Four Mountains Task.

There was no direct instruction for strategy use. There were three practice items followed by 28 test items that were presented in a fixed quasi-random order. Participants were allotted 3 min to complete all test items. Test items varied in the vantage point of the figurine around the display in 45° increments, with every angle presented four times. The shape and color of the objects, the orientation of the layout, and the gender of the figurine were also counterbalanced among items. Four additional items, in which the figurine's and participant's perspectives were the same, were comprehension checks for attentive responding, such that incorrect responses to all four items resulted in exclusion from the analysis. Aside from this, 1PP items were not used in any of the following analyses. In session 2, the same test items were presented in a different fixed order than session 1.

An assumption of many IRT models is that tests are non-speeded (i.e., incorrect responses are due to limited ability rather than not reaching an item in the time allotted; Bolt, Cohen, & Wollack, 2002). Since PTT-A was presented as a timed test, only the first 16 of the items were used in the present analysis, as approximately 90% of the sample reached at least these items in the time allotted and IRT item parameter estimation can be robust to violations of the non-speeded assumption when unreached items are excluded from the analysis (Oshima, 1994). Of note, the total score on the first 16 items of the task was significantly correlated with the total score of the last 16 items ($r = .98$, $p < .05$).

3.3.3. *Four Mountains Test*

A computerized version of the Four Mountains Test (FMT; Hartley & Harlow, 2012; Hartley et al., 2007) displayed a target image of a detailed three-dimensional landscape with four mountains for 10 s (see Fig. 2c). Immediately after the target image, participants were presented with four new images and asked to determine which was the same scene as the target image but from an alternate perspective. Each item had one correct response and three foils. There were no agents in the stimuli, and all images were pictured from a 1PP. There were also no direct instructions for strategy use. Participants were shown three practice items followed by 15 test items in a fixed order with no time limit for response.

3.3.4. *Spatial Orientation Test*

A computerized version of the Spatial Orientation Test (SOT; Friedman, Kohler, Gunalp, Boone, & Hegarty, 2019) displayed a configuration of two-dimensional objects on the left side of the screen and a response circle on the right (see Fig. 2d). At the start of each trial, participants were asked to imagine they were standing at one object in the layout (e.g., the wheel) and facing a second object (e.g., the barrel). Their imagined position was represented by a dot in the center of the response circle, and their imagined line of sight by a solid arrow pointing upwards. They were then asked to imagine pointing towards a third object (e.g., the tree) and to drag the arrow in the response circle to indicate their pointing direction. Although the instructions encouraged a mental self-rotation strategy, no direct instructions were given for strategy use. Participants completed four practice items followed by 12 test items in a random order. The angular distance between the correct response and the participant's response was measured for each trial. Participants were allotted 5 min to complete the entire task, and the 90% completion rate (proportion of participants that responded to at least 11 out of 12 test items) was 0.94.

The multiple-choice format of PTT-A, PPTT, and FMT resulted in dichotomous correct or incorrect responses. Thus, to facilitate a dichotomous IRT analysis that would be comparable across all four PT tasks, we recoded the continuous angular responses on the SOT such that an angular error less than or equal to 30° in either direction was scored as correct and greater than 30° was scored as incorrect. The final score of the test was the proportion of correct responses.

3.4. *Other measures*

3.4.1. *Mental rotation*

A computerized version of the Mental Rotation Test (MRT; Vandenberg & Kuse, 1978 adapted by Peters et al., 1995) displayed a target figure beside four additional figures, consisting of 10 cubes forming a three-dimensional shape. Of these four options, participants were asked to determine which two figures were identical to the target but shown in a different orientation (i.e., rotated about the y-axis). The remaining two figures were foils that were mirrored versions of the target. There were three practice items followed by two blocks of 12 test items presented in a random order with 3 min for response to each block. Participants received 1 point for each correct trial and lost 1 point for each incorrect trial.

Another test of MR ability (Ganis & Kievit, 2015) used cube-figure stimuli adapted from Shepard and Metzler (1971) by adding shading and depth cues to appear more three dimensional. The results of this task are not reported here as they are not relevant for the present findings.

3.4.2. *Spatial working memory*

In a shortened complex span task (symmetry span; Foster et al., 2015) spatial working memory (SWM) items and distractor symmetry-judgement items were presented in alternation. During SWM items, a 4 × 4 grid of white squares was displayed and participants were asked to remember the location of one square that was filled red. During symmetry judgment items, a new array of black and white squares was displayed and participants were asked to judge if it was bilaterally symmetrical. Each block consisted of a varying number of interleaved items (randomly varied between two and seven), and participants were asked to recall the locations of the red squares in the order they were displayed. There were three practice blocks and three test blocks. To limit the possibility of rehearsal of spatial locations during symmetry judgments, participants were allotted a time limit for symmetry judgments during the test that was equal to 2.5 standard deviations more than their average response time for symmetry judgments during practice. A partial span score was used, which consisted of the total number of locations recalled in the correct order.

A simple span task (Position-Span Task; Frick & Möhring, 2016) was used to measure visuo-SWM. The results of this task are not reported here as they are not relevant for the present findings.

3.4.3. *General intelligence*

The Word Reading Subtest of the Wide Range Achievement Test (WRAT; Wilkinson & Robertson, 2006) was used as a measure of verbal fluency/verbal intelligence. Participants were given a list of 55 words of increasing difficulty and asked to pronounce them aloud. The experimenter who was trained on the correct pronunciations ahead of time scored the accuracy of each pronunciation. The total score was the proportion of correctly pronounced words.

Two subtests of the Wechsler Abbreviated Scale of Intelligence, Second Edition, (WASI-II; Wechsler, 2011) were used as a measure of cognitive intelligence. The results from this task are only used in the dropout effects analysis reported in the Results section below.

4. Results

4.1. *Dropout effects*

Means for all measures administered at Session 1 as well as self-reported age and sex were compared for the 48 participants who returned for Session 2 and the 62 who did not. There were no significant differences in mean scores for any of the four PT tasks, MR, SWM, nor the demographic variables (all *t*s < 0.79, all *p*s $> .17$). However, the mean WRAT score was

Table 2
Pearson correlations among original PT tasks with confidence intervals and test–retest reliability

Measure	1	2	3	4	5
1. Photographic Perspective Taking Task (S1)					
2. Photographic Perspective Taking Task (S2)[a]	.80**				
	[.67, .89]				
3. Perspective-Taking Task for Adults (S1)	.69**	.82**			
	[.57, .77]	[.70, .90]			
4. Perspective-Taking Task for Adults (S2)[a]	.66**	.79**	.77**		
	[.46, .80]	[.66, .88]	[.62, .87]		
5. Spatial Orientation Task	.58**	.77**	.61**	.77**	
	[.44, .69]	[.62, .86]	[.48, .72]	[.62, .87]	
6. Four Mountains Task	.36**	.40**	.35**	.47**	.28**
	[.18, .51]	[.12, .61]	[.17, .51]	[.22, .67]	[.09, .44]

Note. Missing data are excluded pairwise;[a]$n = 48$; S1 = Session 1, S2 = Session 2.

significantly higher for participants who returned for Session 2 ($M = 0.87$, $SD = 0.07$) than for those who did not ($M = 0.82$, $SD = 0.10$), $t(107) = -2.847$, $p = .005$, indicating a higher verbal fluency in the return sample. Yet, the two-subtest WASI-II intelligence scores of the return sample were well within the expected range for a typical adult population (Min = 79, Max = 117, $M = 98.23$, $SD = 8.82$), suggesting that this selection process did not greatly reduce generalizability.

4.2. Descriptive statistics

Descriptive statistics for all measures are shown in Table 1. Analyses of sex differences are reported at the end of the Results section (cf. Table 5). Table 2 contains Pearson correlations among all PT measures. Before IRT-informed refinement, FMT showed the weakest correlations with other PT measures (rs < .47), whereas the remaining three tasks all showed moderate correlations with each other (rs > .58).

4.3. Reliability and response times

Internal consistency and general factor saturation were assessed for all four tasks in the form of Cronbach's alpha and McDonald's omega, respectively. Consistency and saturation were high for PPTT at session 1 ($\alpha = .83$; $\omega = .87$) and session 2 ($\alpha = .81$; $\omega = .86$) and for PTT-A at session 1 ($\alpha = .79$; $\omega = .83$) and session 2 ($\alpha = .77$; $\omega = .82$). SOT also showed high consistency ($\alpha = .77$) and saturation ($\omega = .81$). FMT showed only moderate internal consistency ($\alpha = .47$) and saturation ($\omega = .51$). Test–retest reliability analysis was conducted for PPTT and PTT-A. There was a strong positive correlation between session one and session two scores for both PPTT ($r = .82, p < .001$) and PTT-A ($r = .77, p < .001$), indicating good test–retest reliability. As this is the first study in which these measures were administered to adults, we also analyzed response times to assure an appropriate level of challenge. Average response time per item was 5.27 s for PPTT (*SD* = 0.87, Min = 3.46, Max = 7.19) and 8.22 s for PTT-A (*SD* = 2.96, Min = 3.42, Max = 27.24).

4.4. Unidimensionality

To test whether PT tasks measured a unidimensional ability, we first fit a unidimensional IRT model to the data. We assessed the latent structure of the dichotomous data using an approach that combines exploratory factor analysis and Monte Carlo estimation (Drasgow & Parsons,). This test was implemented for each task individually and for all PT task items together in R using the "unidimTest" function of the "ltm" package (Rizopoulos, 2006). Results of this modified parallel analysis (cf. Fig. 3) indicated that the second observed eigenvalue was not substantially larger than the simulated eigenvalue, and therefore the assumption of unidimensionality was met for each task and all items combined.

4.5. Local dependence

Another assumption of unidimensional IRT is that items are not locally dependent (LD). That is, only a person's ability level should determine item responses, and no additional patterns among residuals should remain after this ability is accounted for (Tennant & Conaghan, 2007). Thus, whereas the tests of unidimensionality identified one primary ability captured by each PT task and their combined items, here we checked if there were additional unintended factors that exist beyond this ability. It is important to identify and minimize sources of LD to achieve accuracy in the estimation of item parameters, ability parameters, and information functions (Edelen & Reeve, 2007; Toland, 2014). Thus, we attempted to reduce LD for accuracy in predicting both precision of measurement and individual item effectiveness, while also trying to preserve as many items for inclusion in the unidimensional IRT analysis as possible. The following LD analyses were conducted within-task.

To examine the residual correlations for each item pair, we used the Q_3 statistic, as it has the most detection power among popular LD indexes (Chen & Thissen, 1997; Yen, 1984). This was implemented in R using the "Q3" function of the "sirt" package (Robitzsch, 2019). A critical value of 0.2 above the average residual correlation was chosen based on our sample size and the number of items in each PT task (Christensen, Makransky, & Horton, 2017). Residual correlations revealed that overall, there were three- to five-item pairs per task with a Q_3 statistic above the critical value, suggesting that those items were LD. Next, we followed the suggestions of Toland (2014) and diagnosed items as impactful sources of LD if (a) an item was involved in multiple LD pairs, (b) similarities appeared in stimulus content among LD pairs, (c) item and model fit statistics improved after removal of suspect items, and (d) meaningful differences in estimated slope parameters arose after removal of suspect items. If an item met these criteria, it was removed from the task before IRT analysis to reduce task LD and, in turn, improve the accuracy of parameter estimates.

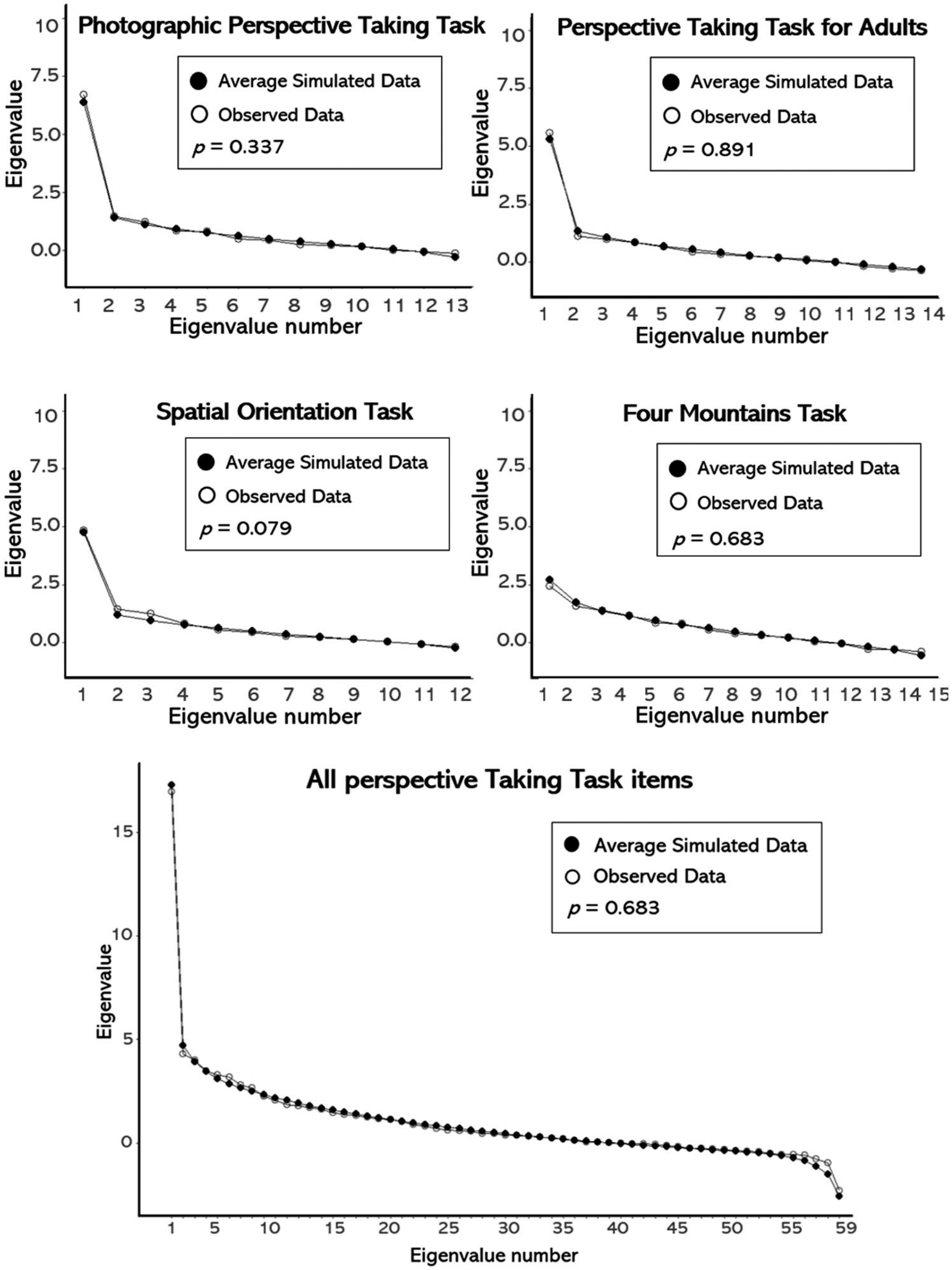

Fig 3. Parallel analysis plots for unidimensionality check for perspective-taking task items.

There were three item pairs within PTT-A with a Q_3 statistic 0.2 or greater above the mean residual correlation, and item 14 appeared among multiple LD pairs; however, removal of item 14 resulted in only minimal differences in item parameter estimates. Thus, there was no indication that including this item was problematic in terms of LD.

Within PPTT, there were two item pairs with a Q_3 statistic 0.2 or greater above the mean residual correlation. A concerning level of LD was present for items 3 and 6, as they had a Q_3 statistic of 0.64, which is well above typically expected values. Sensitivity calibrations indicated that removal of item six eliminated all sources of LD and resulted in good overall and item fit; thus, it was removed from subsequent analyses.

Within SOT, there were four-item pairs with a Q_3 statistic slightly above the critical value. Discernable similarities among their stimuli existed in their starting viewing direction. Although items 4 and 10 had different starting viewing directions, the target objects were perceptually very similar (drum and barrel, respectively). The most concerning source of LD was from items 11 and 12. Their only discernable similarity was that the objects participants were "facing" were both on the left of the object array. An adapted model which excluded either item 11 or 12 eliminated LD in all suspect item pairs and showed good overall and item fit; however, since changes in parameter estimates were miniscule, LD was not deemed problematic.

Finally, among items within FMT with Q_3 values above the critical value, item 15 appeared among three LD pairs. Inspection of the stimuli did not reveal any obvious similarities among these pairs. Sensitivity calibrations with the removal of item 15 showed changes in parameter estimates; however, the LD pair of items 7 and 2 still showed a concerningly large Q_3 value of 0.24. Sensitivity calibrations indicated an adapted model that excluded items 15 and 7 resulted in the greatest changes in parameter estimates, best reduction of overall LD, and good overall and item fit statistics; thus, those items were removed from subsequent analyses.

4.6. Within-task IRT

After establishing that assumptions of unidimensionality and LD were met for items within each task, dichotomous unidimensional 2PL models were fit to the response data. This was implemented in R using the "mirt" package (Chalmers, 2012), and all models were calculated with the Bock–-Aitkin marginal maximum likelihood (EM) estimation method (Bock & Aitkin, 1981). The goodness of fit of the model to each item was assessed using the signed chi-square (S-χ^2) item-fit statistic (Orlando & Thissen, 2000, 2003). A non-significant S-χ^2 value with a root-mean-square error of approximation (RMSEA) close to zero is indicative of good fit.

After removal of one LD item, discriminability of 12 PPTT items ranged from moderate to very high, with a very high average discriminability among items (see Table 3 and Fig. 4a). However, difficulty was on average quite easy, and items only reached a medium level of difficulty. Signed chi-square tests indicated all items had good item fit to the 2PL model (S-χ^2 ps $>$ .05; RMSEAs $<$ 0.10) except for item 4 (S-$\chi^2 = 20.53$, $p < .05$, RMSEA $= 0.15$). The stimulus of item 4 required a 45° angle transformation, which is very close to a 1PP and may explain the lack of fit. After removal of item 4, item information functions (IIFs)

Table 3
Discriminability and difficulty estimates for within-task unidimensional 2PL IRT models

PT Task	Number of Items	Discriminability				Difficulty			
		M	*SD*	Minimum	Maximum	*M*	*SD*	Minimum	Maximum
Original									
SOT	12	1.40	0.72	0.60	2.79	−0.18	0.55	−0.96	0.57
PTT-A	14	1.31	0.46	0.61	2.00	−1.12	0.47	−2.05	−0.36
PPTT	13	2.12	1.67	0.61	6.80	−0.59	0.71	−2.02	0.42
FMT	15	0.66	0.67	−0.09	2.28	−1.87	6.60	−12.94	18.59
Refined Model 1									
SOT	12	1.40	0.72	0.60	2.79	−0.18	0.55	-0.96	0.57
PTT-A	14	1.31	0.46	0.61	2.00	−1.12	0.47	−2.05	−0.36
PPTT	12	1.76	0.62	0.75	3.00	−0.64	0.68	−2.04	0.37
FMT	13	0.59	0.67	−0.24	2.24	−0.53	3.30	−4.32	7.88
Refined Model 2									
SOT	11	1.45	0.77	0.66	2.84	−0.23	0.53	−0.96	0.61
PTT-A	14	1.31	0.46	0.61	2.00	−1.12	0.47	−2.05	−0.36
PPTT	12	1.80	0.62	0.81	2.83	−0.64	0.73	−2.16	0.37
FMT	13	0.89	0.90	0.23	3.13	−1.61	1.03	−3.47	−0.32

Note. Original = parameter estimates for all items excluding first-person perspective comprehension checks. Refined Model 1 = parameter estimates after additionally removing problematic LD items. Refined Model 2 = parameter estimates after additionally removing items with negative or near zero discriminability, near-zero items information, and/or significant signed chi-square value.

showed that all PPTT items captured a good amount of information (area under IIFs > 0.75). The test information function (TIF) indicated that together the 11 best PPTT items captured information across low to moderate ranges of PT ability ($-2 > \theta > 1$), with the most amount of measurement precision for average PT ability (Max. information $\cong 8$; standard error of estimate (SEE) $\cong 0.40$; cf., Fig. 5a).

Parameter estimates of the 14 PTT-A items showed a wide range of discriminability from low to very high, with a moderate average discriminability among items (cf. Table 3). Difficulty of items was easy on average, which is reflected in the locations of item characteristic curves (ICCs) shifted toward the top left of the graph in Fig. 4b. According to signed chi-square tests, all items showed good item fit to the 2PL model ($S\text{-}\chi^2$ ps $> .05$; RMSEAs < 0.10). IIFs showed that most PTT-A items captured a good amount of information (area under IIFs > 0.48). The TIF indicated that together, 14 PTT-A items captured information across a broad range of PT ability ($-4 > \theta > 4$), with the most information collected for moderately low PT ability (cf., Fig. 5b).

ICCs for 12 SOT items can be seen in Fig. 4a. According to signed chi-square tests, all items showed good item fit to the 2PL model ($S\text{-}\chi^2$ ps $> .05$; RMSEAs < 0.10). Discriminability parameters of these items had a wide range from low to very high, with an average discriminability being generally high among items (cf. Table 3). The difficulty of items ranged from very easy to hard with the average being about medium difficulty. Notably, items 6 and 8 had nearly identical ICC parameters and mostly overlapping IIFs (see Fig. 5C). Given the

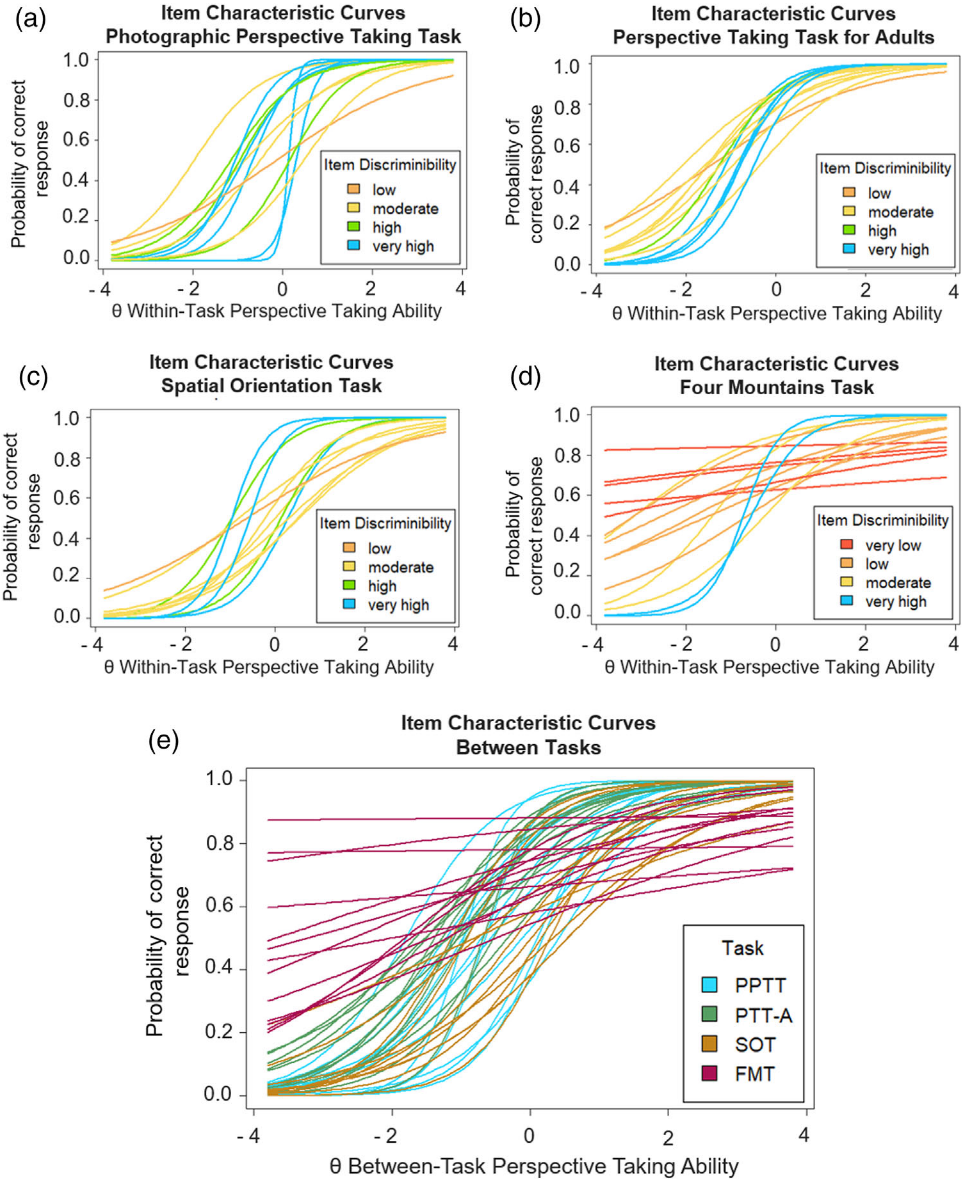

Fig 4. Item characteristic curves for original perspective-taking tasks. Abbreviations: PPTT, Photographic Perspective Taking Task; PTT-A, Perspective Taking Task for Adults; SOT, Spatial Orientation Task; FMT, Four Mountains Task.

redundancy of these items and the small amount of item information each contributes to the test (area under IIF < 1), one item was removed from the battery to improve task efficiency with negligible sacrifice to the amount of total information measured within the task. The TIF for the 11-item SOT indicated that overall, measurement was most precise for a small range of average within-task ability from -1.5 to .5 (Max. information $\cong 6$; SEE $\cong 2.5$). There was

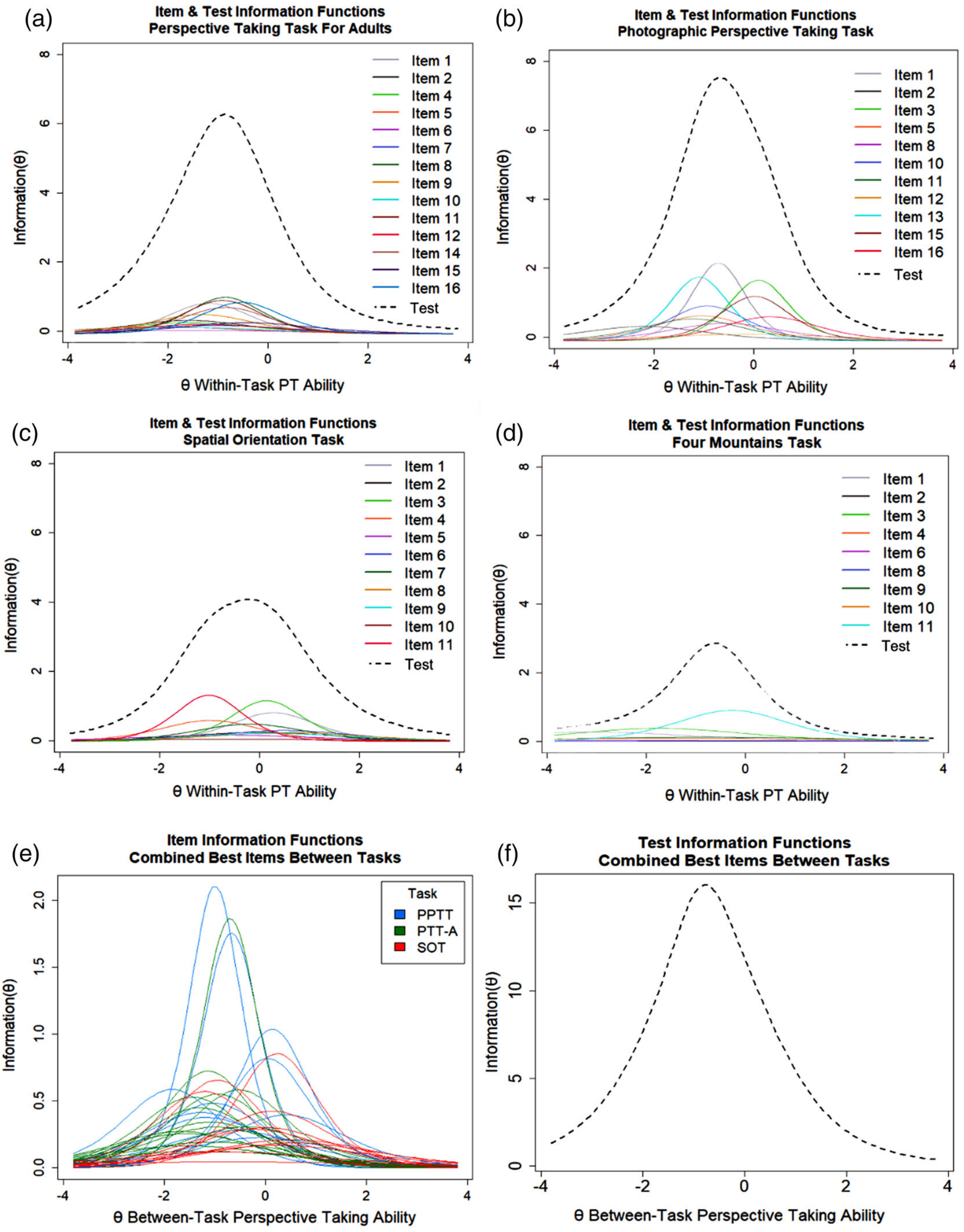

Fig 5. Item and test information functions for perspective-taking tasks. (a) Item and test inforamtion for all items of Perspective-Taking Task for Adults. (b) Item 6 which was a source of local dependence (LD) and item 4 which did not show sufficient item fit to the model are not plotted. (c) Item 12 was a source of LD and not plotted. (d) Items 5 and 7 were sources of LD and were not plotted. (e) Item information from best items of PTT-A, PPTT, and SOT together. (f) Test information for best items of PTT-A, PPTT, and SOT together.

a small amount of information (Max. information $\cong 4$; SEE $\cong 0.5$) for high ($0.5 < \theta < 2$) and low ($-2.5 < \theta < -1.5$) within-task abilities, and less than 0.2 information at extreme ranges of ability.

After removing two LD items from the original FMT, the discriminability of 13 FMT items ranged from none to very high, with low being the average level of discriminability among items (cf. Table 3). Items 2, 4, 6, and 8 had slopes very close to zero and/or negative values, showing little discrimination among individuals of different ability levels (cf. Fig. 4d). Thus, these items seemed to assess an ability which differs from the other items within the measure and were removed. On average, the items had medium difficulty and ranged from very easy to very hard. Signed chi-square tests indicated all items had good item fit to the 2PL model (S-χ^2 *p*s > .05; RMSEAs < 0.08). IIFs showed that most FMT items had little measurement precision (average area under IIFs = 0.53; cf. Fig. 5d).

4.7. Between-tasks IRT

Item response data from the refined tasks were fit to (A) a unidimensional 2PL IRT model, (B) a non-hierarchical between-item MIRT model, and (C) a hierarchical bifactor MIRT model (see Fig. 1). As a metric of model fit, the limited-information M_2 statistic was calculated (Maydeu-Olivares & Joe, 2006) with its accompanying *p* value and RMSEA using quasi-Monte Carlo integration. A lower non-significant M_2 value with an RMSEA close to zero is indicative of a well-fitting model (De Ayala, 2009; Toland, 2014). Although we do report the comparative fit index (CFI) of Pearson's χ^2 test statistic, M_2 was used as the primary assessment of model fit as it is more robust to Type I errors when modeling a large number of dichotomous items in both unidimensional 2PL and multidimensional models (Xu, Paek, & Xia, 2017). Comparison of non-nested models was conducted using Vuong's test (Schneider, Chalmers, Debelak, & Merkle, 2019; Vuong, 1989) and implemented via the "nonnest2" R package (Merkle, You, Schneider, Bae, & Merkle, 2018).

First, Model A did not show a good fit to the data when items from all tasks were combined ($M_2 = 1235$, $p < .05$; RMSEA = 0.02, CFI = 0.96) as indicated by the large and significant M_2 value. We investigated ICCs to compare item discriminability and difficulty in terms of a between-task PT ability. As can be seen in Fig. 4E, items from the FMT showed ICCs with shallower slopes in comparison to the ICCs of all three other tasks, indicating that FMT is measuring a different ability than the other tasks. For this reason, we reran Model A without FMT. Although the M_2 statistic decreased, the value still remained significant suggesting the unidimensional 2PL IRT model still did not show a good fit to the data ($M_2 = 690$, $p < .05$; RMSEA = 0.04, CFI = 0.97).

Next, we tested the fit of Model B and Model C, maintaining the exclusion of FMT. To assess whether the PT tasks measured distinct but correlated abilities, we tested the fit of Model B. Results indicated that Model B was also not a good fit to the data ($M_2 = 761$, $p < .01$; RMSEA = 0.05, CFI = 0.98). Finally, by fitting the data to Model C, we tested the alternative possibility that PPTT, PTT-A, and SOT measure a common PT ability, but also have additional uncorrelated dimensions associated with each task which contribute to individual differences in responses. Model C showed excellent fit to the data as indicated by having the lowest and only non-significant M_2 value among all competing models ($M_2 = 595$, $p = .136$; RMSEA = 0.02, CFI = 0.99).

According to Vuong's test of model fit, Model A fit the data significantly better than the Model B ($z = 2.75$, $p < .01$), but not better than Model C ($z = -3.63$, $p < .01$). A final implementation of Vuong's test indicated that, indeed, Model C showed a better fit to the data than Model A ($z = -6.05$, $p < .001$). Thus, the best fit was found for Model C (Fig. 1C), which assumes that PPTT, PTT-A, and SOT all measure a common underlying PT ability, but also have additional uncorrelated dimensions within measures that contribute to individual differences in PT responses.

5. Divergent validity of PT

To test for the extent to which MR scores are related to PT scores after controlling for sex, general intelligence, and SWM, we conducted two stepwise hierarchical regression analyses, each consisting of two steps. In Step 1, all control variables were entered. In Step 2, participants' MR performance was added to derive the percentage of unique variance explained by MR-related effects. In the first regression, the dependent PT variable was a composite PT score calculated by averaging the z-scores of all four PT tasks in their original forms. In the second regression, the dependent PT variable was the theta estimate (i.e., ability estimate) of the general PT factor Model C (Fig. 1C).

Results of the first hierarchical regression are presented in pane a in Table 4. After controlling for sex, general intelligence, and SWM, 9% of the variance in pre-IRT PT composite scores was explained by the MR measure. Results of the second hierarchical regression, are presented in pane b in Table 4. After controlling for sex, general intelligence, and SWM, MR explained 6% of the variance in PT general factor theta estimates.

To test if a hybrid task comprised of the best items across PTT-A, PPTT, and SOT tasks showed greater divergent validity from MR than any one task alone, we conducted partial correlations of each task and the hybrid with MR, removing the effects of G and SWM. PTT-A showed the greatest amount of divergent validity from MR as indicated by a small partial correlation value ($r = .23$, $p < .05$), followed by PPTT ($r = .34$, $p < .01$). The hybrid measure had a larger partial correlation than either of these two tasks ($r = .43$, $p < .01$). Finally, SOT had the largest partial correlation value ($r = .45$, $p < .01$), showing the least amount of divergent validity from MR.

5.1. Sex differences

Independent samples t tests were conducted for performance on all measures. As shown in Table 5, males had significantly higher scores than females on all PT tasks (ps $< .05$, ds > 0.5). Males also had significantly higher scores than females on MRT, $t(109) = 3.38$, $p < .001$, $d = 0.88$. There were no statistically significant differences in mean scores on SWM, $t(109) = 1.88$, $p = .063$, nor for general intelligence, $t(109) = 1.82$, $p = .071$.

Table 4
Stepwise regressions predicting PT

	Step 1			Step 2		
Variable	*B*	*SE*	*p*	*B*	*SE*	*p*
(a) Stepwise regression predicting original PT measures as a composite score.						
Intercept	−2.67	0.62	<0.01	−2.97	0.58	0.00
Female	−0.60	0.15	<0.01	−0.31	0.16	0.05
G	2.50	0.73	<0.01	2.17	0.68	0.00
SWM	0.04	0.01	<0.01	0.03	0.01	0.00
MRT		-		0.06	0.01	0.00
R^2		0.387			0.474	
F		23.26**			24.88**	
ΔR^2		-			0.087	
ΔF		-			1.62**	
(b) Stepwise regression predicting theta estimates from general PT factor of the bifactor MIRT model						
Intercept	−3.00	0.75	0.00	−3.33	0.71	0.00
Female	−0.68	0.17	0.00	−0.39	0.18	0.03
G	2.81	0.87	0.00	2.51	0.83	0.00
SWM	0.04	0.01	0.00	0.04	0.01	0.00
MRT		-		0.06	0.02	0.00
R^2		0.372			0.436	
F		22.55**			22.03**	
ΔR^2		-			0.064	
ΔF		-			−0.52**	

Note. ** $p < .001$.Abbreviations: G, general intelligence, Wide Range Achievement Test – Reading Subtest; SWM, spatial working memory, Symmetry Span; MRT, mental rotation Test; Original PT composite score, the average of the *z*-score of each task in its original form.

6. Discussion

This study used IRT to examine the psychometric properties of four PT measures, assess their divergent validity from MR, and to probe the need for and feasibility of a composite measure that validly and reliably assesses a wide range of PT ability. Accurate measurement of PT ability will have important applications in education, navigation, and individual differences research and will serve to inform theories on differences and similarities between spatial skills. In regard to reliability, we found that PTT-A and PPTT, which were originally designed for developmental research but were adapted for use in adults, showed good test–retest reliability. In addition, PTT-A, PPTT, and SOT all showed good internal consistency and construct saturation.

Next, we found that each task on its own fit well to 2PL unidimensional IRT models, suggesting they each measure one uniform ability within-task. Further, within each task, there were at least a few items that did not discriminate among PT abilities or did not contribute a great deal to the overall amount of information being collected, and thus, could be removed to decrease test-administering time and increase estimation accuracy. Also, we excluded 1PP

Table 5
Sex differences in performance

	Male		Female		*t*-Test [a]	*d*	CI
	M	*SD*	*M*	*SD*			
Perspective-Taking Tasks							
Spatial Orientation Task	0.75	0.2	0.49	0.2	4.76**	1.1	[0.62, 1.58]
Four Mountains Task	0.78	0.1	0.7	0.2	2.52*	0.58	[0.12, 1.05]
Photographic Perspective Taking Task (S1)	0.82	0.2	0.65	0.2	3.20**	0.74	[0.27, 1.21]
Perspective Taking Task for Adults (S1)	0.64	0.2	0.51	0.2	3.13*	0.72	[0.26, 1.19]
Mental Rotation Tests							
MRT	13.3	4.5	7.73	4.4	5.50**	1.27	[0.78, 1.76]
Spatial working memory							
Symmetry Span	30.2	7.9	27	7.4	1.83		
General intelligence							
WRAT	0.87	0.1	0.84	0.1	1.76		

Note: *$p < .05$, **$p < .01$. Equality of variance was not assumed;Abbreviations: S1, session 1; MRT, mental rotation Test; WRAT, Wide Range Achievement Test- Reading.

items in our IRT analysis; however, researchers are encouraged to still include these items in the tasks as attention checks or baseline for reaction time.

After within-task IRT-refinement, we assessed convergent validity among tasks. We found that a combination of all items across tasks captured a great deal of information from a broad range of PT ability, except for FMT which seemed to measure a different ability than the other three. Removing FMT items improved convergent validity among tasks. FMT is often used to assess spatial memory in early diagnosis of Alzheimer's disease (Chan et al., 2016). Thus, the paradigm stands out from the other three tasks in that spatial memory is heavily taxed. Future work should carefully consider if there are any other aspects of task paradigms that interfere with the measurement of PT performance.

In terms of dimensionality, a hierarchical bifactor model (Fig. 1c) best fit the data, suggesting that secondary factors influenced scores in addition to the first general PT factor. One such secondary factor that may further explain individual differences on PT tasks is the presence of agents in task stimuli. Work from Tarampi, Heydari, and Hegarty (2016) showed that female but not male participants performed significantly better on PT tasks when agents were present in the stimuli. Indeed, we found a larger effect size for sex differences on SOT, which had no agents, than for PTT-A and PPTT, which included agents. Previous work has found that female participants use embodied strategies to perform PT tasks more often than male participants (Kaiser et al., 2008; Kessler & Wang, 2012). Thus, differences in performance may not be due to lack of ability, but rather to individual differences in strategy choice (i.e., Hegarty et al., 2018) in combination with the nature of the task stimuli. Future studies should further elucidate how other task features may impact PT performance in the face of alternate strategy use.

In addition, future work might also consider investigating the variation of strategy use on specific PT tasks and how these may impact performance. In the present study, we approached the question of whether participants really use PT while taking different tasks with a data-driven approach by assessing discriminant validity from MR. However, introspective approaches (e.g., think-aloud protocols) may also be useful and appropriate in different contexts (Barratt, 1953).

When relating PT scores to MR, we found that even after controlling for sex, general intelligence, and SWM, there was still 9% shared variance in performance between these two spatial skills; however, when deriving a PT ability score from the most promising items as informed by a bifactor MIRT model, MR only shared 6% variance with PT performance. This indicates that the IRT-refined composite PT measure minimized sources of unintentional tasks variance and thus exhibited better discriminant validity. However, we have not fully exhausted the psychometric possibilities here. Future studies should continue to assess psychometric properties of tasks for precise measurement of individual differences in spatial abilities.

In regard to creation of a hybrid measure of PT, our results did not diagnose an immediate need for such a combined measure under most circumstances. First, although there were variations in the amount of item information gathered across ranges of ability for each task, there was a large overlap regarding the ability ranges they assessed. Therefore, combining the measures would offer generally more test information, but not necessarily measure a wider range of PT ability than any one task alone. Second, a hybrid task did not show greater divergent validity from MR than most of the individual tasks. In particular, SOT showed less divergent validity from MR than PTT-A and PPTT, and combining these tasks into a hybrid led to a decrease in divergent validity, making it more associated with MR than PTT-A or PPTT alone. Third, we did not find evidence that SOT, PTT-A, and PPTT measured distinct abilities (Model B); rather, Model C showed the best fit to the data, suggesting that the tasks measured a common PT ability, and any one task may be used in place of another.

Some limitations of the present study should be considered when interpreting these results. Three of the four PT tasks in the battery were timed either at the test or item level, with PTT-A being considered a heavily timed task; thus, while we used the recommended approaches for IRT analysis under timed conditions, parameter estimates may be moderately impacted when considering the full scale of 32 items under timed pressure. The number of items completed in the allotted time may also be an indicator of proficiency and yield additional information in the high-ability range.

In addition, the sample size of the current study may be considered modest in comparison to other IRT studies; however, there are several factors that are considered when selecting sample size in IRT analysis including the type of items (i.e., dichotomous vs. graded response), the number of items, and the number of parameters being estimated (see Sahin & Anil, 2017). According to Morizot, Ainsworth, and Reise (2007) as cited in Thorpe and Favia (2012, pp. 21–22), “an unbiased analysis for dichotomously-scored items (those with two possible response codes, e.g., 0 or 1) may have as few as 100 participants.” Thus, our current sample size of 110 seems to be sufficient for retaining unbiased estimates of dichotomously scored items in the models we have presented. However, future studies should take advantage of larger sample sizes to estimate more complex models, such as 3PL IRT models.

Finally, previous research indicates that women tend to have relatively low spatial ability scores relative to men (Voyer, Voyer, & Bryden, 1995). Although the sample-independent assumption of IRT should allow for broader generalizability of results beyond the demographics of the used sample, it should still be taken into consideration that the present study was largely made up of female participants.

To conclude, there are several implications of this study for the accurate and reliable measurement of PT. First, the fit of the bifactor model (Model C) to data from PTT-A, PPTT, and SOT support the idea that tasks which measure PT are multidimensional to some extent. That is, particular stimulus features of PT tasks recruit a common PT ability, but also conceptually distinct abilities respective to each task that additionally impact individual differences in scores. However, creating a truly process-pure task is difficult. Thus, researchers can leverage a bifactor model approach to derive more precise scores (e.g., Carr et al., 2020) for both the primary PT dimension and the subdomains separately. Second, for situations in which IRT is not feasible, the psychometric characteristics of each task reported here indicated that PTT-A, PPTT, or SOT may be used on their own as robust, reliable, and valid measures of PT ability. Finally, IRT-refinement at the item and task level resulted in an increase of divergent validity of PT from MR. In the future, IRT approaches, and in particular a bifactor model approach (e.g., Reise, 2012), can continue to be used to provide an accurate description of individual differences data in other domains.

Acknowledgments

This work was supported by grant 1660996 from the NSF Division of Research on Learning.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Aichhorn, M., Perner, J., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Do visual perspective tasks need theory of mind?. *Neuroimage*, *30*(3), 1059–1068.

Amorim, M.-A., & Stucchi, N. (1997). Viewer- and object-centered mental explorations of an imagined environment are not equivalent. *Cognitive Brain Research*, *5*(3), 229–239. https://doi.org/10.1016/S0926-6410(96)00073-0

Barratt, E. S. (1953). An analysis of verbal reports of solving spatial problems as an aid in defining spatial factors. *The Journal of Psychology*, *36*(1), 17–25.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348.

Burgess, N., Spiers, H. J., & Paleologou, E. (2004). Orientational manoeuvres in the dark: Dissociating allocentric and egocentric influences on spatial memory. *Cognition*, *94*(2), 149–166.

Burles, F., Slone, E., & Iaria, G. (2017). Dorso-medial and ventro-lateral functional specialization of the human retrosplenial complex in spatial updating and orienting. *Brain Structure and Function*, *222*(3), 1481–1493.

Carr, M., Horan, E., Alexeev, N., Barned, N., Wang, L., & Otumfuor, B. (2020). A longitudinal study of spatial skills and number sense development in elementary school children. *Journal of Educational Psychology*, *112*(1), 53.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Chan, D., Gallaher, L. M., Moodley, K., Minati, L., Burgess, N., & Hartley, T. (2016). The 4 mountains test: A short test of spatial memory with high sensitivity for the diagnosis of pre-dementia Alzheimer's disease. *Journal of Visualized Experiments*, *11*(116), e54454.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American psychologist*, *12*(11), 671.

Clements-Stephens, A. M., Vasiljevic, K., Murray, A. J., & Shelton, A. L. (2013). The role of potential agents in making spatial perspective taking social. *Frontiers in Human Neuroscience*, *7*, 497. https://doi.org/10.3389/fnhum.2013.00497

Committeri, G., Galati, G., Paradis, A. L., Pizzamiglio, L., Berthoz, A., & LeBihan, D. (2004). Reference frames for spatial cognition: Different brain areas are involved in viewer-, object-, and landmark-centered judgments about object location. *Journal of Cognitive Neuroscience*, *16*(9), 1517–1535.

Cooper, L. A., & Regan, D. T. (1982). Attention, perception, and intelligence. In R. J. Sternberg (Ed.), *Handbook of Human Intelligence* (pp. 123–169). New York, NY: Cambridge University Press.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2020). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology: General*, *150*(2), 242–275. https://doi.org/10.1037/xge0000783

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, *68*(3), 363.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(1), 5.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Fields, A. W., & Shelton, A. L. (2006). Individual skill differences and large-scale environmental learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 506–515. https://doi.org/10.1037/0278-7393.32.3.506

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378.

Flavell, J. H., Flavell, E. R., Green, F. L., & Wilcox, S. A. (1981). The development of three spatial perspective-taking rules. *Child Development*, *52*, 356–358. https://doi.org/10.2307/1129250

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*(2), 226–236. https://doi.org/10.3758/s13421-014-0461-7

Frick, A. (2019). Spatial transformation abilities and their relation to later mathematics performance. *Psychological Research*, *83*(7), 1465–1484. https://doi.org/10.1007/s00426-018-1008-5

Frick, A., & Möhring, W. (2016). A matter of balance: Motor control is related to children's spatial and proportional reasoning skills. *Frontiers in Psychology*, *6*, 2049. https://doi.org/10.3389/fpsyg.2015.02049

Frick, A., Möhring, W., & Newcombe, N. S. (2014). Picturing perspectives: Development of perspective-taking abilities in 4- to 8-year-olds. *Frontiers in Psychology*, *5*, 386. https://doi.org/10.3389/fpsyg.2014.00386

Friedman, A., Kohler, B., Gunalp, P., Boone, A. P., & Hegarty, M. (2019). A computerized spatial orientation test. *Behavior Research Methods*, *52*, 799–812. https://doi.org/10.3758/s13428-019-01277-3

Ganis, G., & Kievit, R. (2015). A new set of three-dimensional shapes for investigating mental rotation processes: Validation data and stimulus set. *Journal of Open Psychology Data*, *3*, e3. http://doi.org/10.5334/jopd.ai

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436. https://doi.org/10.1007/BF02295430

Gibbons, R. D., Rush, A. J., & Immekus, J. C. (2009). On the psychometric validity of the domains of the PDSQ: An illustration of the bi-factor item response theory model. *Journal of Psychiatric Research*, *43*(4), 401–410. https://doi.org/10.1016/j.jpsychires.2008.04.013

Hartley, T., Bird, C. M., Chan, D., Cipolotti, L., Husain, M., Vargha-Khadem, F., & Burgess, N. (2007). The hippocampus is required for short-term topographical memory in humans. *Hippocampus*, *17*(1), 34–48. https://doi.org/10.1002/hipo.20240

Hartley, T., & Harlow, R. (2012). An association between human hippocampal volume and topographical memory in healthy young adults. *Frontiers in Human Neuroscience*, *6*, 338. https://doi.org/10.3389/fnhum.2012.00338

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, *32*(2), 175–191. https://doi.org/10.1016/j.intell.2003.12.001

Hegarty, M. (2018). Ability and sex differences in spatial thinking: What does the mental rotation test really measure?. *Psychonomic Bulletin & Review*, *25*(3), 1212–1219.

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166–184.

Huttenlocher, J., & Presson, C. C. (1973). Mental rotation and the perspective problem. *Cognitive Psychology*, *4*(2), 277–299. https://doi.org/10.1016/0010-0285(73)90015-7

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, *40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Kaiser, S., Walther, S., Nennig, E., Kronmüller, K., Mundt, C., Weisbrod, M., … Vogeley, K. (2008). Gender-specific strategy use and neural correlates in a spatial perspective taking task. *Neuropsychologia*, *46*(10), 2524–2531.

Kessler, K., & Wang, H. (2012). Spatial perspective taking is an embodied process, but not for everyone in the same way: Differences predicted by sex and social skills score. *Spatial Cognition & Computation*, *12*(2-3), 133–158. https://doi.org/10.1080/13875868.2011.634533

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, *22*(2), 154–169.

Kozhevnikov, M., & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition*, *29*(5), 745–756. https://doi.org/10.3758/BF03200477

Kozhevnikov, M., Motes, M. A., Rasch, B., & Blajenkova, O. (2006). Perspective-taking vs. mental rotation transformations and how they predict spatial navigation performance. *Applied Cognitive Psychology*, *20*(3), 397–417. https://doi.org/10.1002/acp.1192

Lambrey, S., Doeller, C., Berthoz, A., & Burgess, N. (2012). Imagining being somewhere else: Neural basis of changing perspective in space. *Cerebral Cortex*, *22*(1), 166–174. https://doi.org/10.1093/cercor/bhr101

Lempers, J. D., Flavell, E. R., & Flavell, J. H. (1977). The development in very young children of tacit knowledge concerning visual perception. *Genetic Psychology Monographs*, *95*(1), 3–53.

Liu, Y., Magnus, B., O'Connor, H., & Thissen, D. (2018). Multidimensional item response theory. In P. Irwing, D. Hughes, & T. Booth (Eds.), *The Wiley handbook of psychometric testing* (pp. 445–493). Hoboken, NJ: Wiley-Blackwell. https://doi.org/10.1002/9781118489772.ch16

Lohman, D. F. (1996). Spatial ability and g. *Human Abilities: Their Nature and Measurement*, *97*(116), 1.

Lohman, D. F. (1979). *Spatial ability: A review and reanalysis of the correlational literature*. Technical report. School of Education, Stanford University, Stanford, CA.

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. Charlotte, NC: Information Age Publishing, Reading, MA.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, *1*(1), 1–11.

Masangkay, Z. S., McCluskey, K. A., McIntyre, C. W., Sims-Knight, J., Vaughn, B. E., & Flavell, J. H. (1974). The early development of inferences about the visual percepts of others. *Child Development*, *45*(2), 357–366.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713.

Merkle, E., You, D., Schneider, L., Bae, S., & Merkle, M. E. (2018). Package 'nonnest2.' *Psychological Methods*, *21*, 151–163.

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 407–421).

Nazareth, A., Newcombe, N. S., Shipley, T. F., Velazquez, M., & Weisberg, S. M. (2019). Beyond small-scale spatial skills: Navigation skills and geoscience education. *Cognitive Research: Principles and Implications*, *4*(1), 17. https://doi.org/10.1186/s41235-019-0167-2

Nazareth, A., Weisberg, S. M., Margulis, K., & Newcombe, N. S. (2018). Charting the development of cognitive mapping. *Journal of Experimental Child Psychology*, *170*, 86–106. https://doi.org/10.1016/j.jecp.2018.01.009

Newcombe, N. S., & Shipley, T. F. (2015). Thinking about spatial thinking: New typology, new assessments. In J. S. Gero (Ed.), *Studying visual and spatial reasoning for design creativity* (pp. 179–192). Dordecht, the Netherlands: Springer Netherlands. https://doi.org/10.1007/978-94-017-9297-4_10

Oldakowski, R. K. (2001). Activities to develop a spatial perspective among students in introductory geography courses. *Journal of Geography*, *100*(6), 243–250. https://doi.org/10.1080/00221340108978451

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*(4), 289–298.

Oshima, T. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*(3), 200–219.

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test—Different versions and factors that affect performance. *Brain and Cognition*, *28*(1), 39–58. https://doi.org/10.1006/brcg.1995.1032

Piaget, J., & Inhelder, B. (1956). *The child's conception of space*. London: Routledge & Kegan Paul.

Pittalis, M., & Christou, C. (2010). Types of reasoning in 3D geometry thinking and their relation with spatial ability. *Educational Studies in Mathematics*, *75*(2), 191–212.

Plummer, J. D., Bower, C. A., & Liben, L. S. (2016). The role of perspective taking in how children connect reference frames when explaining astronomical phenomena. *International Journal of Science Education*, *38*(3), 345–365. https://doi.org/10.1080/09500693.2016.1140921

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks pædagogiske Institut.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*(1), 19–31. https://doi.org/10.1007/s11136-007-9183-7

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

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

Robitzsch, A. (2019). sirt: Supplementary item response theory models. R package version 3.9-4. https://CRAN.R-project.org/package=sirt

Sahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Journal of Educational Sciences: Theory & Practice*, *17*(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2014). A task is a task is a task: Putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in Psychology*, *5*, 1475. https://doi.org/10.3389/fpsyg.2014.01475

Schneider, L., Chalmers, R. P., Debelak, R., & Merkle, E. C. (2019). Model selection of nested and non-nested item response models using Vuong tests. *Multivariate Behavioral Research*, *55*(5), 664–684.

Shelton, A. L., Clements-Stephens, A. M., Lam, W. Y., Pak, D. M., & Murray, A. J. (2012). Should social savvy equal good spatial skills? The interaction of social skills with spatial perspective taking. *Journal of Experimental Psychology: General*, *141*(2), 199–205. https://doi.org/10.1037/a0024617

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703. https://doi.org/10.1126/science.171.3972.701

Tarampi, M. R., Heydari, N., & Hegarty, M. (2016). A tale of two types of perspective taking: Sex differences in spatial ability. *Psychological Science*, *27*(11), 1507–1516.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research*, *57*(8), 1358–1362.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577. https://doi.org/10.1007/BF02295596

Thorpe, G. L., & Favia, A. (2012). Data Analysis using item response theory methodology: An introduction to selected programs and applications.

Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, *34*(1), 120–151. SAGE Publications. https://doi.org/10.1177/0272431613511332

Uttal, D. H., & Cohen, C. A. (2012). Spatial thinking and STEM education: When, why, and how? In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. *57*, pp, 148–153). New York, NY: Academic Press. https://doi.org/10.1016/B978-0-12-394293-7.00004-2

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, *47*(2), 599–604. https://doi.org/10.2466/pms.1978.47.2.599

Vogeley, K., May, M., Ritzl, A., Falkai, P., Zilles, K., & Fink, G. R. (2004). Neural correlates of first-person perspective as one constituent of human self-consciousness. *Journal of Cognitive Neuroscience*, *16*(5), 817–827.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*(2), 250–270. https://doi.org/10.1037/0033-2909.117.2.250

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*(2), 307–333. https://doi.org/10.2307/1912557

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, *101*(4), 817–835. https://doi.org/10.1037/a0016127

Wechsler, D. (2011). *WASI-II: Wechsler abbreviated scale of intelligence*. San Antonio, TX: PsychCorp.

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide range achievement test (WRAT4)*. Lutz, FL: Psychological Assessment Resources.

Wraga, M., Creem, S. H., & Proffitt, D. R. (2000). Updating displays after imagined object and viewer rotations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 151–168. https://doi.org/10.1037/0278-7393.26.1.151

Xu, J., Paek, I., & Xia, Y. (2017). Investigating the Behaviors of M 2 and RMSEA2 in Fitting a Unidimensional Model to Multidimensional Data. *Applied Psychological Measurement*, *41*(8), 632–644. http://doi.org/10.1177/0146621617710464

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145.

Zacks, J. M., Mires, J., Tversky, B., & Hazeltine, E. (2000). Mental spatial transformations of objects and perspective. *Spatial Cognition and Computation*, *2*(4), 315–332. https://doi.org/10.1023/A:1015584100204