Journal of Experimental Child Psychology 160 (2017) 50–66

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

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

Testing the validity of a continuous false belief task in 3- to 7-year-old children

Caitlin E.V. Mahy [a,*], Daniel M. Bernstein [b], Lindsey D. Gerrard [a], Cristina M. Atance [c]

[a] Department of Psychology, Brock University, St. Catharines, Ontario L2S 3A1, Canada
[b] Department of Psychology, Kwantlen Polytechnic University, Surrey, British Columbia V3W 2M8, Canada
[c] School of Psychology, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada

ARTICLE INFO

Article history:
Received 1 June 2016
Revised 13 March 2017

Keywords:
False belief
Theory of mind
Inhibition
Early childhood
Validity
Continuous measurement

ABSTRACT

In two studies, we examined young children's performance on the paper-and-pencil version of the Sandbox task, a continuous measure of false belief, and its relations with other false belief and inhibition tasks. In Study 1, 96 children aged 3 to 7 years completed three false belief tasks (Sandbox, Unexpected Contents, and Appearance/Reality) and two inhibition tasks (Head–Shoulders–Knees–Toes and Grass/Snow). Results revealed that false belief bias—a measure of egocentrism—on the Sandbox task correlated with age but not with the Unexpected Contents or Appearance/Reality task or with measures of inhibition after controlling for age. In Study 2, 90 3- to 7-year-olds completed five false belief tasks (Sandbox, Unexpected Contents, Appearance/Reality, Change of Location, and a second-order false belief task), two inhibition tasks (Simon Says and Grass/Snow), and a receptive vocabulary task (Peabody Picture Vocabulary Test). Results showed that false belief bias on the Sandbox task correlated negatively with age and with the Change of Location task but not with the other false belief or inhibition tasks after controlling for age and receptive vocabulary. The Sandbox task shows promise as an age-sensitive measure of false belief performance during early childhood and shows convergent and discriminant validity.

* Corresponding author.
E-mail address: caitlin.mahy@brocku.ca (C.E.V. Mahy).

http://dx.doi.org/10.1016/j.jecp.2017.03.010
0022-0965/

Introduction

For the past 40 years, young children's theory of mind (ToM) has been predominantly measured by standard false belief tasks (Gopnik & Astington, 1988; Premack & Woodruff, 1978; Wellman, 1990; Wellman, Cross, & Watson, 2001; Wellman & Liu, 2004; Wimmer & Perner, 1983). In a standard false belief task, such as Change of Location, children learn about two characters who initially share knowledge about the location of a ball in a box. The first character then leaves the room, and while she is gone a second character moves the ball to the basket. When the first character returns, children are typically asked two types of questions, namely, where will the first character look for the ball? (in the box or the basket; false belief question) and where is the ball really? (memory control question). To pass the false belief question, children must appreciate that the first character holds a false belief about the item's location because she did not observe the movement of the ball by the second character. Children must also inhibit their own knowledge of the item's true location to pass the false belief question. Finally, children's responses are scored as pass or fail depending on whether they can answer both the false belief and memory control questions (Baron-Cohen, Leslie, & Frith, 1985; Wimmer & Perner, 1983).

Researchers have suggested moving away from these pass/fail false belief tasks (Birch & Bloom, 2007; Bloom & German, 2000) because (a) they require abilities other than ToM, such as inhibitory control, working memory, and language (e.g., Carlson & Moses, 2001; German & Leslie, 2000; Milligan, Astington, & Dack, 2007; Riggs, Peterson, Robinson, & Mitchell, 1998; Roth & Leslie, 1998), and (b) ToM involves abilities beyond understanding false beliefs, such as emotion understanding, the ability to imitate intended and completed actions, modifying one's behavior based on others' knowledge states, and detecting agency in non-animate objects that move as if they were animate (e.g., Carpenter, Akhtar, & Tomasello, 1998; Johnson, Slaughter, & Carey, 1998; O'Neill, 1996). The field, however, has persisted in the use of such tasks. These false belief tasks are able to detect age-related changes in false belief understanding between 3 and 5 years of age; however, they are of limited utility beyond 5 years once children can pass such tasks. Accordingly, the need for new ToM tasks has been identified, and some new measures have been developed (Begeer, Bernstein, van Wijhe, Scheeren, & Koot, 2012; Bloom & German, 2000; Devine & Hughes, 2013; Dumontheil, Apperly, & Blakemore, 2010; Lagattuta, Sayfan, & Harvey, 2013; Peterson, Wellman, & Slaughter, 2012; Sommerville, Bernstein, & Meltzoff, 2013; Tahiroglu et al., 2014).

For example, a continuous measure of preschoolers' false belief understanding that diverges from the typical pass/fail scoring of standard false belief tasks has recently been introduced. The "Sandbox task" yields a continuous score that indexes children's false belief bias (Begeer et al., 2012; Sommerville et al., 2013). This score measures how biased children are by the *true* location of an object when they reason about a person's *false belief* about the location of the object. In the real-object version of the Sandbox task, children are placed in front of a three-dimensional box and hear a story about a protagonist who puts an object in one location (L1). Then, while the protagonist is absent, a second character moves the object to a second location (L2). Children are then asked the critical false belief question of where the first protagonist will look for the object when he returns. To answer this question, children point to where the protagonist will look for the object in the sandbox rather than selecting one of two options of where the protagonist will look as is typical in standard false belief tasks (e.g., in the box or in the basket). Based on the difference between their response and the location where the first protagonist should look (L1), children receive a false belief bias score (measured in centimeters or millimeters). This continuous measure of false belief has important advantages as compared with a dichotomous choice because it does not treat false belief understanding as an all-or-none phenomenon. Furthermore, the task has the potential to detect more subtle development of false belief understanding when comparing children under and over 5 years of age.

Begeer and colleagues (2012) and Coburn, Bernstein, and Begeer (2015) developed a paper-and-pencil version of the Sandbox task in which children (aged 6 years and older) and adults viewed a picture of the sandbox (Fig. 1) rather than a three-dimensional object. In this version, an "X" was marked to show where the protagonist placed the object and then where the second character moved the object while the protagonist was gone (these marks remained visible during the story). Children then

"Judy and her dad are planting flowers in the planter box to surprise her mom. Judy's dad buries the flower here (*experimenter points to X at 0 mm*) and then goes to the shed to find a shovel."

0 mm

"While Judy's dad is gone, Judy decides to move the flower here (*experimenter points to X at 100 mm*)"

100 mm

"When Judy's dad comes back, where is he going to look for the flower? Show me where he will look by marking the spot with this pen."

Fig. 1. Example Sandbox story and images in the false belief trial in Study 1.

indicated where the protagonist would look for the object when he returned by marking an X on a blank sandbox (where the Xs were no longer visible). The paper-and-pencil version of the Sandbox task has not been administered to children younger than 6 years.

Similar to the standard Change of Location task, the Sandbox task requires that children represent the protagonist's false belief about the location of the object and inhibit their knowledge of its actual location to respond correctly. In contrast to the standard Change of Location task, however, the Sandbox task measures children's responses continuously by examining the distance in centimeters or millimeters between where children have searched or marked an X and the object's original location (L1; where the protagonist *should* look based on his false belief about the object's location). This distance is referred to as the false belief bias score and represents how much children's knowledge of the object's current location biases their judgment of where the protagonist will look. Longer distances from the original location (L1) represent worse false belief understanding because they show that the current location of the object (L2) biases children's decision.

An important feature of many false belief tasks is sensitivity to age-related changes in ToM development. Performance on standard false belief tasks increases rapidly between 3 and 5 years of age and often shows a sharp increase indicating a qualitative shift in performance at around 4 or 4.5 years when representational ToM develops (i.e., children come to represent mental states and understand that these states often differ from reality). This is the age at which children typically begin to pass these tasks (Wellman et al., 2001). Although the real-object Sandbox task was sensitive to age differences in false belief understanding between young children (3- and 5-year-olds) and adults on long distance trials (Sommerville et al., 2013), the paper-and-pencil version used with older children and adolescents (6–20 years) who were typically developing or had an autism spectrum disorder did not reveal age-related differences in false belief understanding (Begeer et al., 2012). Begeer et al. (2012) failure to detect age differences may have been due to a smaller response range on the paper-and-pencil version. This interpretation is plausible in light of the finding that short distance trials in Sommerville et al. (2013) real-object version revealed no performance differences between 3- and 5-year-olds on shorter distance trials. Alternatively, it is possible that the paper-and-pencil version of the Sandbox task is not sensitive enough to capture the more subtle developments in false

belief understanding in the 6- to 20-year age range. Thus, it is important to examine the paper-and-pencil version of this task in a wider age range of young children to assess its utility in capturing developments in false belief understanding. A goal of the current study, therefore, was to examine whether the paper-and-pencil Sandbox task is feasible to administer to younger children (3- to 7-year-olds) and, importantly, whether performance would be sensitive to developmental increases in false belief understanding given its relative ease of administration compared with the real-object version.

Relations between performance on the Sandbox task and other false belief tasks

Although performance on the real-object Sandbox task correlated modestly with a Change of Location task for 3- and 5-year olds previously (Sommerville et al., 2013), to the best of our knowledge nothing has been published about how it relates to other standard false belief tasks such as the Unexpected Contents and Appearance/Reality tasks. This is important because the standard Change of Location and Sandbox tasks share a nearly identical structure. For example, characters see the original location of the object, only one character is privileged in seeing the object move, and then children are asked about the naive character's belief about the object's location. Thus, an interesting question is whether children's performance on the Sandbox task relates to false belief tasks that *differ* in their structure. For example, the Unexpected Contents and Appearance/Reality tasks do not involve an object moving locations but instead involve updating a previous incorrect belief about the contents of a box or the identity of an item (Flavell, Flavell, & Green, 1983). In the Unexpected Contents task, children may see a Band-Aids box containing crayons. Children are asked about their own previous belief and another naive child's belief about the contents of the box in a forced-choice format (e.g., crayons or Band-Aids; Gopnik & Astington, 1988). Similarly, in the Appearance/Reality task, children see a sponge that resembles a rock. After learning its true nature, children indicate what they thought the sponge was when they first saw it (self question), a rock or a sponge, and then indicate what a naive child who did not see the sponge squeezed would think the identity of the item is, a rock or a sponge. The Unexpected Contents and Appearance/Reality tasks both require children to reflect on their previous naive knowledge state and another child's current naive knowledge state.

An open question is whether the Sandbox task (which asks children only about another character's mental state) will relate more strongly to responses about another child's knowledge state in the Unexpected Contents and Appearance/Reality tasks compared with responses about children's own previous mental states. Notably, performance on the Unexpected Contents and Appearance/Reality tasks correlate positively with performance on the standard Change of Location task (e.g., Carlson & Moses, 2001). With the greater variance in continuous responses on the Sandbox task, it is an unanswered question whether Unexpected Contents and Appearance/Reality task performance will correlate positively with Sandbox task performance.

Relations with inhibitory control

In typical standard false belief tasks, children have two discrete response options (e.g., Band-Aids *or* crayons in the Unexpected Contents task). Inhibitory control, in particular, has been found to correlate positively with false belief tasks (Carlson & Moses, 2001)—likely due to the inhibitory demand to avoid the option that corresponds with reality (where the item really is what the true content or identity is) to choose the option that corresponds with a false belief. It is still unknown whether children's difficulty with the inhibitory demands of the task is at a conceptual level (the conflict between reality and false belief more generally) or at the response selection level (selecting the correct option while inhibiting selection of the item's true location). The Sandbox task has the potential to address this question because it requires children to inhibit their knowledge of the object's current location to reason from the perspective of a naive character (i.e., conceptually, it requires inhibition), but arguably it has less inhibitory demand at the response selection level because children are not required to make a forced-choice response. In contrast, most standard false belief tasks require *both* conceptual inhibition and response selection level inhibition. With respect to the latter, children must choose between one salient response that represents the true state of the world and another less salient, but correct, response that represents a false belief. Given that standard false belief tasks correlate highly with inhi-

bitory control, a goal of the current study was to examine relations between the Sandbox task and inhibitory control. To our knowledge, no research has examined the relation between performance on the Sandbox task and measures of inhibitory control in children, although Bernstein, Thornton, and Sommerville (2011) found that inhibition correlated positively with false belief bias on the Sandbox task in adults.

Study 1

The main purpose of this study was to examine whether performance on the paper-and-pencil version of the Sandbox task in younger children (3- to 7-year-olds) would (a) be sensitive to age-related changes (and possible to administer to young preschoolers), (b) relate to two other false belief tasks with different task structures (Unexpected Contents and Appearance/Reality) and with self and other components, and (c) correlate with measures of inhibitory control. An overarching goal was to establish construct validity of the paper-and-pencil Sandbox task in a wider age range of young children because ToM research often lacks sufficient psychometrics to establish validity (convergent and discriminant validity in particular).

Method

Participants

A total of 96 children (51 girls and 45 boys) participated: 20 3-year-olds (11 girls; M_{age} = 42.80 months, SD = 3.09), 21 4-year-olds (12 girls; M_{age} = 52.71 months, SD = 3.94), 18 5-year-olds (9 girls; M_{age} = 67.00 months, SD = 3.01), 19 6-year-olds (9 girls; M_{age} = 78.32 months, SD = 3.28), and 18 7-year-olds (9 girls; M_{age} = 88.72 months, SD = 3.69). We based this sample size on power calculations that to detect medium effect sizes (r = .35, alpha = .05, power = .95) we would need at least 83 children. Children were mostly Caucasian (85%) and from middle-class backgrounds (80% of families had an annual income of more than $40,000). Children were recruited from university developmental databases and were tested at two sites in North America.

Measures

Theory of mind. Unexpected Contents task. In the Unexpected Contents task (Gopnik & Astington, 1988), children were shown a closed Band-Aids box and asked what they thought was inside. The majority of children answered Band-Aids (n = 90), and the remaining 6 children readily agreed with the experimenter when he or she suggested that there might be Band-Aids inside the box. Then, the experimenter opened the box to reveal crayons. Once the Band-Aids box was closed, the experimenter asked children what they thought was inside the box before it was opened (self question) and then what another child about their age would think was inside the box if he or she came into the room when the box was closed (other question): crayons or Band-Aids. The self/other questions and the response options (Band-Aids/crayons) were counterbalanced across participants. Finally, to confirm that children remembered the true contents of the Band-Aids box (crayons), they were asked to report what was really inside the box. Only data from children who passed this memory control question were included in the analyses (n = 87).

Appearance/Reality task. In the Appearance/Reality task (Flavell et al., 1983), children were shown a sponge that looked like a rock inside a transparent plastic box and were asked what they thought it was. Most children reported that it was a rock (n = 86), and the remaining 10 children agreed with the experimenter when he or she suggested it looked like a rock. Then, the experimenter removed the sponge from the plastic box, squeezed it, and allowed children to touch it, revealing that it was a sponge. Once the sponge was put back inside the box, children were asked what they thought it was before they touched it (self question) and what another child would think it was if they came into the room right then (other question): a rock or a sponge. The order of presentation for the self and other questions and response options (rock/sponge) were counterbalanced across participants so that children were randomly assigned to one of two counterbalanced groups. To ensure that all children remembered that the object was a sponge, children were asked what the object really was: a rock

or a sponge. Only data from children who passed this memory control question were included in the analyses (n = 89).

Sandbox task. In the Sandbox task (adapted from Begeer et al. (2012)), children were introduced to two characters: Judy and her father. Children were told the following story: "Judy and her father are planting flowers in the planter box to surprise her mother. Judy's father buries a flower here [marked with an X that was 25 mm from the edge of the first planter box], and then he goes to the shed to get a shovel. While Judy's father is gone, Judy moves the flower here [marked with an X that was 100 mm left of the first X on the second planter box]." Then, these planter boxes were removed and children were asked, "When Judy's father comes back, where will he look for the flower?" Children were asked to mark an X on a blank planter box to indicate where Judy's father would look for the flower (Fig. 1). Proportional bias scores were calculated based on the horizontal distance between the child's X and the first X (in millimeters) where Judy's father would look, if children understood the existence of a false belief, divided by 100 mm (this was to permit visual comparisons with the results of Study 2). Positive proportional bias scores (close to or greater than 1) represent greater false belief bias, whereas scores closer to zero (or slightly negative) represent less false belief bias.

Inhibitory control. Head–Shoulders–Knees–Toes task. In the Head–Shoulders–Knees–Toes task (Ponitz et al., 2008), after a warm-up phase where children were asked to follow the experimenter's commands, children were asked to do the opposite of what the experimenter said; they were told to touch their head when the experimenter told them to touch their toes and to touch their toes when the experimenter told them to touch their head. Children were given four practice trials and needed to get at least three of the four trials correct to advance to the test trials (otherwise their data were excluded). A total of 19 children failed to meet criterion on these practice trials and, thus, were excluded from the analysis. Then, 10 Head–Toes (HT) test trials were administered where children needed to do the opposite of what the experimenter said and demonstrated. Children were given a score out of 20 on the HT trials based on whether they responded incorrectly (0 points), needed to correct themselves (1 point), or responded correctly (2 points). After the HT trials, children were introduced to a new rule: to touch their shoulders when the experimenter told them to touch their knees and to touch their knees when the experimenter told them to touch their shoulders. Four practice trials were administered in which children needed to demonstrate basic understanding of the rules by succeeding on at least three of the four trials. Then, children began the second phase where they completed 10 Head–Shoulders–Knees–Toes (HSKT) trials that included all four types of instructions: touch their head, touch their toes, touch their shoulders, and touch their knees. Children needed to respond by performing the opposite response and were given a score on these HSKT trials out of 20. Children's scores on the HT and HSKT trials were combined to yield a total score out of a possible 40 points. Interrater reliability on 34% of the data was substantial (Cohen's kappa = .76; Landis & Koch, 1977).

Grass/Snow task. In the Grass/Snow task (Carlson & Moses, 2001), children were asked to report the color of snow and grass to the experimenter. All children responded correctly. The experimenter then introduced the rules to a "silly opposites" game where children needed to point to the white square when the experimenter said "grass" and to the green square when the experimenter said "snow." Children completed 2 practice trials in which they needed to point to the correct color square and then place their hands back on handprints in front of the color squares. Once children passed these two trials, 14 additional trials were administered for a total of 16 trials. Children received a score out of 32 based on whether each trial was correct (2 points), self-corrected (1 point), or incorrect (0 points). Interrater reliability on 34% of the data was substantial (Cohen's kappa = .71; Landis & Koch, 1977).

Procedure

Children completed the tasks in a fixed order (see Carlson & Moses, 2001): Unexpected Contents, Appearance/Reality, Sandbox, HSKT, and Grass/Snow. Parents completed basic demographics information while their children participated in the study. The procedure took approximately 30 min in total. The research ethics boards at the University of Oregon and the University of Ottawa approved all procedures.

Results

Means and standard deviations for performance on all tasks by age appear in Table 1.

Relations with age

Table 2 shows correlations among age and performance on all tasks. Age in months correlated negatively with the false belief bias score on the Sandbox task, $r(91) = -.304$, $p = .003$, indicating that younger children marked their X closer to the true location of the bulb, whereas older children marked their X closer to the previous location of the bulb (the correct place that Judy's father would look for the bulb; Fig. 2).[1] Of note in Fig. 2 is the fact that children of all ages used the entire continuous space of the Sandbox task to respond rather than responding bimodally, indicating that children treated the task continuously rather than categorically. For the standard false belief tasks, age in months correlated positively with Unexpected Contents false belief (self and other questions), $rs(86) > .509$, $ps < .001$, and Appearance/Reality (self and other questions), $rs(89) > .400$, $ps < .001$. Age also correlated positively with the two measures of inhibition, HSKT and Grass/Snow, $rs(77) > .373$, $ps < .001$.

Relations between Sandbox task and standard false belief tasks

Table 2 shows point-biserial correlations among false belief tasks as well as correlations after controlling for age in months. False belief bias score on the Sandbox task correlated negatively with Unexpected Contents for other, $r(83) = -.325$, $p = .003$, Appearance/Reality (for self), $r(85) = -.271$, $p = .012$, and Appearance/Reality (for other), $r(85) = -.252$, $p = .020$. After controlling for children's age in months, however, correlations between false belief bias scores and Unexpected Contents and Appearance/Reality disappeared.

Relations with inhibitory control

False belief bias score on the Sandbox task did not correlate with HSKT and Grass/Snow performance, $rs(68) > -.111$, $ps > .311$ (see Table 2). This was not due to lack of correlations between standard false belief tasks and these inhibition measures more generally given that performance on the Unexpected Contents and Appearance/Reality tasks (self and other) correlated positively with performance on HSKT, $rs(73) > .377$, $ps < .002$, and Unexpected Contents and Appearance/Reality (self) were positively correlated with Grass/Snow, $rs(87) > .292$, $ps < .006$. Interestingly, Unexpected Contents and Appearance/Reality (other) performance was not significantly related to Grass/Snow performance, $rs(87) < .183$, $ps > .090$. After we controlled for age, HSKT (but not Grass/Snow) performance remained correlated with Appearance/Reality (self) performance, $r(70) = .46$, $p < .001$.

Discussion

Study 1 had three main results. First, the paper-and-pencil version of the Sandbox task has sufficient ability to detect age-related improvements in false belief understanding in young children. Second, the Sandbox task was unrelated to standard false belief measures with different structures from the Change of Location task after controlling for age (this was true for both self and other components). Third, the Sandbox task did not correlate with inhibitory control.

After we controlled for age-related variance, bias on a single false belief trial of the Sandbox task failed to correlate with the Unexpected Contents and Appearance/Reality tasks, or our two measures of inhibition. It is possible that false belief bias failed to correlate with these two measures of false belief because their structure differs from that of the typical Change of Location task after which the Sandbox task is modeled (demonstrating discriminant validity). Whereas the Unexpected Contents and Appearance/Reality tasks focus on a knowledge change surrounding an object's identity or contents, the Change of Location and Sandbox tasks focus on a change in an object's physical loca-

[1] To control for the possibility that these correlations were being driven solely by 3- to 5-year-old children, we re-ran our analyses on preschoolers only. There was no significant relation between age in months and egocentric bias on the Sandbox task in 3- to 5-year-olds. This null effect indicates that the correlation between age and false belief bias score on the Sandbox task was not due solely to preschoolers and that the added power provided by combining all ages drove the effect.

Table 1
Means and standard deviations on all tasks by age group in Study 1.

	3-year-olds	4-year-olds	5-year-olds	6-year-olds	7-year-olds
Unexpected Contents (self)	.14 (.36)	.50 (.51)	.71 (.47)	.79 (.42)	.94 (.25)
Unexpected Contents (other)	.23 (.44)	.60 (.50)	.59 (.51)	.90 (.32)	.94 (.24)
Appearance/Reality (self)	.21 (.43)	.43 (.51)	.83 (.38)	.95 (.23)	.88 (.33)
Appearance/Reality (other)	.43 (.51)	.62 (.50)	.67 (.49)	.95 (.23)	.88 (.33)
Proportional false belief bias on Sandbox task	.78 (.61)	.56 (.67)	.71 (.64)	.38 (.61)	.26 (.40)
Head–Shoulders–Knees–Toes	10.88 (10.43)	19.76 (11.85)	29.67 (5.89)	29.84 (5.59)	31.11 (7.69)
Grass/Snow	20.00 (10.97)	25.89 (5.41)	29.41 (2.18)	26.95 (3.87)	28.44 (2.73)

Note. Standard deviations are in parentheses. Ranges for scores on task are as follows: Unexpected Contents (self) (0–1), Unexpected Contents (other) (0–1), Appearance/Reality (self) (0–1), Appearance/Reality (other) (0–1), proportional false belief bias on Sandbox task (−.25 to 2.0; higher scores denote worse performance), Head–Shoulders–Knees–Toes (0–40), Grass/Snow (0–32).

Table 2
Correlations among age, theory of mind measures, and inhibition for Study 1.

	2	3	4	5	6	7	8
1. Age in months	−.30**	.53**	.51**	.55**	.40**	.61**	.37**
2. Sandbox false belief bias		−.21† (−.10)	−.33** (−.21†)	−.27* (−.11)	−.25* (−.14)	−.21 (.02)	−.11 (.03)
3. Unexpected Contents (self)			.46** (.27*)	.65** (.53**)	.28* (.09)	.41** (.14)	.33* (.18)
4. Unexpected Contents (other)				.44** (.25*)	.57** (.46**)	.42** (.21†)	.18† (.02)
5. Appearance/Reality (self)					.37** (.19†)	.61** (.45**)	.29* (.14)
6. Appearance/Reality (other)						.38** (.20†)	.17 (.12)
7. Head–Shoulders–Knees–Toes							.19 (.12)
8. Grass/Snow							

Note. Partial correlations controlling for age are in parentheses. Degrees of freedom range from 68 to 91.
† $p < .10$.
* $p < .05$.
** $p < .01$.

tion. Further, it seems that inhibition does not correlate with false belief bias scores after we controlled for age, whereas performance on the HSKT task still correlated with performance on most of the false belief tasks after we controlled for age. These results support the idea that the Sandbox task might impose less inhibitory demand at the response selection level. These results are notable because they suggest some discriminant validity between the Sandbox task and standard measures of false belief with a different structure that measures children's understanding of their own and others' previous false beliefs. This suggests that not all false belief tasks are equal; children likely show more similar performance on tasks that share a common structure. Further, this suggests that the Sandbox task might not share the inhibitory demand at the response level that most standard false belief measures contain.

Nevertheless, a limitation of the current study was the use of a single false belief trial of the Sandbox task and a lack of memory control questions. We administered only one false belief trial on the Sandbox task. Best practices suggest that multiple trials as well as memory questions would result in a more thorough assessment of children's performance on the task (see Sommerville et al., 2013). Also important to note is that several children in the current study marked an X outside of the space between L1 and L2. It is difficult to know whether these responses represented an exceptional egocentric bias, a lack of understanding the task, or children's difficulty in remembering where the X was or having poor motor control to mark a more accurate location for L2. Thus, including mem-

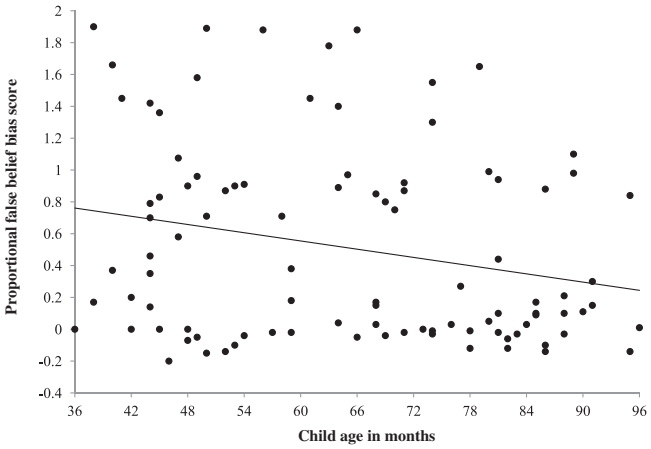

Fig. 2. Mean false belief bias scores on the Sandbox task and children's age in months in Study 1. R^2 = .09. Note that false belief bias scores of 1 reflect biased responding, whereas scores of 0 reflect perfect accuracy and unbiased responding.

ory control trials will allow us to control for children's accuracy when no false belief reasoning is present and to control for poor memory or motor control while children mark the X. Another limitation was the existence of ceiling effects for 7-year-olds in particular on our standard measures of false belief; however, these children's performance nevertheless contributed to important variance because the correlations between Sandbox task performance and age in the current study were not driven by 3- to 5-year olds' performance. An ongoing challenge in the field is to design ToM measures that capture variance across a wide age range, which the Sandbox task seems to do well. We designed Study 2 to address Study 1's limitations.

Study 2

In Study 2, we attempted to replicate the findings of Study 1 and also make several improvements to its design. Given that we administered only one false belief trial of the Sandbox task in Study 1, we aimed to replicate our findings with more trials while controlling for task-specific memory. Accordingly, Study 2 included four trials of the Sandbox task (two false belief and two memory control) so that we could control for children's memory bias when examining their false belief bias. As with the standard change of location task, the Sandbox task contains memory control trials to ensure that participants remember and follow the stories. In addition to the standard false belief measures used in Study 1—Unexpected Contents and Appearance/Reality—we included the Change of Location task in Study 2 to replicate previous correlations between the Change of Location task and false belief bias scores on the Sandbox task (Sommerville et al., 2013). We also included a second-order false belief task for 5- to 7-year-olds old to avoid ceiling levels of performance on the standard false belief tasks in the oldest age groups (as suggested by Study 1 results) and to further examine conceptual distinctions between first- and second-order false belief tasks. Because most 3-year-olds struggled to pass the practice phase of the HSKT task in Study 1, here we used the Simon Says task to minimize data loss from the youngest participants (3-year-olds). Finally, in Study 2 we examined relations among ToM and inhibition tasks after controlling for age and verbal ability by measuring children's receptive vocabulary using the Peabody Picture Vocabulary Test.

Method

Participants

A total of 97 children (54 girls and 43 boys) participated. Of these children, 7 were excluded from the final sample due to refusal to cooperate or difficulty in following instructions (5 3-year-olds and 1

4-year-old) or the presence of a learning disability (1 6-year-old). The final sample consisted of 90 children: 18 3-year-olds (11 girls; M_{age} = 42.06 months, SD = 3.93), 18 4-year-olds (8 girls; M_{age} = 52.22 months, SD = 3.26), 18 5-year-olds (8 girls; M_{age} = 65.56 months, SD = 2.71), 18 6-year-olds (13 girls; M_{age} = 78.67 months, SD = 3.18), and 18 7-year-olds (10 girls; M_{age} = 90.06 months, SD = 17.76). Children were mostly Caucasian (87.8%) and from middle-class backgrounds (75.3% of families had an annual income of more than $40,000), consistent with the population from which they were sampled. Children were recruited from a developmental database at Brock University as well as from local child care centers and preschools.

Measures

For the purpose of brevity, here we describe only additional tasks or tasks that were substantially changed from Study 1.

Sandbox task. The administration of the Sandbox task (adapted from Begeer et al. (2012)) was similar to Study 1 except that instead of using one false belief trial, we administered four trials consisting of two false belief and two memory control trials (see Table 3 for a complete description of the trials). For each trial, children were introduced to two characters and told a story that involved an object being moved from a first location (L1) to a second location (L2). In the false belief trials, children were asked to indicate where a naive character would look for an item that had been moved in their absence. In memory control trials, children were asked where the original location of the item was. As in Study 1, children needed to mark a point in a box to indicate where the character would look. Rather than having children make an X, which proved to be difficult for younger children in Study 1, they were asked to make a dot with a fine-tipped marker. Children's responses were measured in millimeters from the correct location to compute a bias score (horizontal distance) divided by the distance between the first and second locations (to control for different distances between the first and second locations). The trials were administered in a fixed order: memory control (L2 153 mm to the left of L1), false belief (L2 121 mm to the right of L1), memory control (L2 121 mm to the right of L1), and false belief (L2 153 mm to the left of L1) following Coburn et al. (2015) procedure. We then averaged scores across the two false belief trials and also across the two memory control trials.

Table 3
Four trials of the Sandbox Task administered in Study 2.

Trial	Trial type	First scenario	Second scenario	Test question
1	Memory control	"Yoko and her mom have just come home from the grocery store and are putting the groceries away. Yoko puts away the ice cream in the freezer here (L1) and then goes outside to play."	"While Yoko is outside, her mom opens the freezer and moves the ice cream here (L2)."	"Then Yoko comes back inside. Where did she put the ice cream before she went out to play?" **(L1)**
2	False belief	"Judy and her dad are planting flowers in the planter box to surprise her mom. Judy's dad buries a flower here (L1) and then goes to the shed to find a shovel."	"While Judy's dad is gone, Judy put the flower here (L2)."	"When Judy's dad comes back, where is he going to look for the flower?" **(L1)**
3	Memory control	"Sally and Ann are outside playing in the sandbox. Sally hides a toy dog in the sand here (L1) and then goes to get a drink of water."	"While Sally is inside the house, Ann finds the toy dog and hides it here (L2)."	"Then Sally comes back. Where did she put the toy dog before she went inside?" **(L1)**
4	False belief	"Jenny and Billy are walking in the forest. Jenny leaves her backpack near a tree here (L1) and then goes near the stream to cool down."	"While Jenny is gone, Billy moves her backpack here (L2)."	"When Jenny comes back, where will she look for her backpack?" **(L1)**

Note. The correct location is denoted in bold after the test question. Children who placed the dot or X closer to the correct location would have smaller false belief bias scores.

Change of Location task. In this false belief task, children were introduced to Jane, who put her marble in a chest and then went outside to play. While Jane was gone, children were introduced to Sara, who moved the marble from the chest to the box. Then children were asked the critical question of where Jane would look for her marble when she returned: the chest or the box? Children received a score of 1 if they indicated the chest and a score of 0 if they indicated the box. To ensure intact memory for the item's true location, we included only children who answered the memory control question of "where is the marble really?" ($n = 90$; all children passed). Children were always asked the false belief question first and the memory control question second.

Second-order false belief task. In the second-order false belief task (adapted from Perner and Wimmer (1985)), the experimenter used pictures and props to enact a story involving two characters. Children watched a story about John and Mary, who were playing in the park. Mary wanted to buy ice cream but had left her money at home, so she decided to go home and return to the park to get ice cream later that afternoon. Once Mary was gone, John saw the ice cream man leaving the park and asked where he was going. The ice cream man said he was going to sell ice cream at the school. On the way to the school, the ice cream man went by Mary's house and Mary learned that the ice cream man was on his way to the school. Later that afternoon, John went to Mary's house, but Mary's mother told John that Mary had gone to get ice cream. Children were reminded that John didn't know that Mary had talked to the ice cream man. The test question was, "Where does John think Mary has gone to get ice cream?" Children were asked to justify their responses and were scored on whether they referred to John's mental states. Finally, children answered three memory control questions to ensure that they understood and remembered the story correctly (e.g., where did Mary go for her ice cream?; does Mary know that the ice cream man is at the school?; does John know that the ice cream man has talked to Mary?). Only children who answered all three memory control questions correctly were included in the analysis ($n = 48$ of 54 children aged 5–7 years). Children were given a second-order false belief score from 0 to 2 depending on whether they answered the test question and gave an appropriate justification (referring to John's mental state).

Simon Says task. The Simon Says task (Carlson, 2005) was used as a measure of conflict inhibition. First, a female experimenter explained to children that they would be playing a silly game where they should perform an action only if she prefaced the command with "Simon says." Children were told to remain perfectly still otherwise. Children completed two practice trials where they were asked to show the experimenter what they would do if she said, "Simon says clap your hands" and "Clap your hands". If children performed either of the practice trials incorrectly, the experimenter told them what the correct response should be. The experimenter then issued commands in quick succession while demonstrating the actions. Children completed 10 trials (5 with and 5 without "Simon says"). Performance on non-Simon-says trials was taken as an index of inhibitory ability (0 = commanded movement, 1 = partial movement, 2 = different movement, 3 = no movement; scored individually for each non-Simon-says trial; range = 0–15; Carlson & Meltzoff, 2008). Interrater agreement on 28% of non-Simon-says trials was high (Cohen's kappa = .90).

Peabody Picture Vocabulary Test. Children completed the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 2007) as a measure of verbal intelligence. Children were shown four pictures and were asked to point to the picture that corresponded to the word that the experimenter read aloud. Children completed two practice trials where they were given feedback and told that it was okay to guess if they did not know what the word meant. Then children were given the first set of words depending on their chronological age. The test was administered until participants failed 8 of 12 items in a given set. Raw scores on this measure were used in the analysis.

Procedure

Children completed the tasks in a fixed order (see Carlson & Moses, 2001): Sandbox task, Appearance/Reality task, Grass/Snow task, Unexpected Contents task, Simon Says task, Change of Location task, second-order false belief task (if they were 5 years or older), and Peabody Picture Vocabulary Test. Parents completed basic demographics information while their children participated in the

study. The procedure required approximately 45 min to complete and was approved by the Brock University research ethics board.

Results

To control for children's memory in their false belief performance on the Sandbox task, memory bias was controlled for using partial correlations when computing relations among false belief bias scores and other variables of interest.

Relations with age

Table 4 shows performance on all tasks by age, and Table 5 shows correlations among tasks and correlations after we controlled for age in months. Children's age in months correlated negatively with the false belief bias score after controlling for memory bias, $r(87) = -.365$, $p < .001$ (Fig. 3). All of the standard false belief tasks (Unexpected Contents [self and other], Appearance/Reality [self and other], and Change of Location) positively correlated with age in months, $rs(82) > .471$, $ps < .001$. In addition, second-order false belief task performance positively correlated with age in months, $r(48) = .500$, $p < .001$. The two measures of inhibition, Simon Says and Grass/Snow, positively correlated with age, $rs(83) > .616$, $ps < .001$, as well as PPVT performance, $r(88) = .886$, $p < .001$.

Relations among Sandbox, standard false belief, and second-order false belief tasks controlling for age and vocabulary

After controlling for children's age in months, PPVT scores, and memory bias, only the Change of Location task remained significantly negatively correlated with false belief bias score, $r(83) = -.22$, $p = .047$. Performance on the false belief trials of the Sandbox task was uncorrelated with performance on the second-order false belief task after controlling for memory bias, $r(42) = .21$, $p = .169$. Performance on the standard false belief measures (Unexpected Contents, Appearance/Reality, and Change of Location tasks) was also uncorrelated with second-order false belief task performance, $rs(42) < .11$, $ps > .468$.

Relations with inhibitory control

After we controlled for age, PPVT scores, and memory bias, false belief bias did not correlate with measures of inhibition. Similar to Study 1, standard false belief task performance correlated positively with measures of inhibition, $rs(78) > .350$, $ps < .003$. Even after we controlled for age and PPVT scores,

Table 4
Means and standard deviations on all tasks by age group in Study 2.

	3-year-olds	4-year-olds	5-year-olds	6-year-olds	7-year-olds
Proportional false belief bias on Sandbox task	.52 (.36)	.58 (.48)	.26 (.34)	.07 (.39)	.09 (.40)
Proportional memory bias on Sandbox task	.36 (.33)	.35 (.44)	.09 (.33)	.07 (.38)	−.03 (.18)
Unexpected Contents (self)	.53 (.52)	.57 (.51)	.83 (.38)	1.00 (.00)	1.00 (.00)
Unexpected Contents (other)	.47 (.52)	.57 (.51)	.89 (.32)	.94 (.24)	1.00 (.00)
Appearance/Reality (self)	.36 (.50)	.40 (.51)	.82 (.39)	.94 (.24)	1.00 (.00)
Appearance/Reality (other)	.29 (.47)	.40 (.51)	.82 (.39)	.94 (.24)	1.00 (.00)
Change of Location	.22 (.43)	.33 (.49)	.44 (.51)	.89 (.32)	1.00 (.00)
Second-order false belief	–	–	0.20 (0.41)	0.67 (0.82)	1.17 (0.79)
Simon Says	0.40 (1.06)	2.47 (3.60)	10.47 (4.88)	12.06 (3.92)	13.89 (2.08)
Grass/Snow	16.21 (9.88)	22.24 (7.38)	26.61 (7.64)	30.17 (1.62)	30.17 (2.15)
Peabody Picture Vocabulary Test (raw score)	70.78 (21.93)	84.12 (18.50)	108.12 (12.64)	125.83 (14.55)	145.72 (12.77)

Note. Standard deviations are in parentheses. Ranges for scores on task are as follows: Unexpected Contents (self) (0–1), Unexpected Contents (other) (0–1), Appearance/Reality (self) (0–1), Appearance/Reality (other) (0–1), Change of Location (0–1), proportional false belief or memory bias on Sandbox task (−.38 to 1.28), second-order false belief task (0–2), Simon Says (0–15), Grass/Snow (0–32), Peabody Picture Vocabulary Test (raw score).

Table 5
Correlations among age, theory of mind measures, and inhibition for Study 2.

	2	3	4	5	6	7	8	9	10	11
1. Age in months	−.37**	.47**	.52**	.59**	.63**	.62**	.50**	.82**	.62**	.89**
2. False belief bias controlling for memory bias		−.17 (−.04)	−.28* (−.14)	−.22* (−.02)	−.29* (−.07)	−.37** (−.22*)	.07 (.21)	−.33* (−.08)	−.34* (−.18)	−.28* (.08)
3. Unexpected Contents (self)			.46** (.32**)	.70** (.59**)	.38** (.07)	.47** (.30**)	.23 (.05)	.47** (.18)	.50** (.28*)	.41** (.01)
4. Unexpected Contents (other)				.62** (.49**)	.53** (.30*)	.53** (.33**)	.12 (−.07)	.56** (.29*)	.53** (.27*)	.48** (.03)
5. Appearance/Reality (self)					.54** (.19†)	.57** (.35**)	.18 (−.01)	.69** (.46**)	.50** (.20†)	.63** (.36**)
6. Appearance/Reality (other)						.52** (.17)	.28† (.11)	.63** (.21†)	.35* (−.10)	.64** (.20†)
7. Change of Location							.27† (−.07)	.50** (.03)	.41** (.08)	.56** (.04)
8. Second-order false belief								.18 (.01)	.21 (.02)	.48** (.19)
9. Simon Says									.67** (.30*)	.77** (.17)
10. Grass/Snow										.61** (.20†)
11. PPVT										

Note. Partial correlations controlling for age in months and PPVT scores are in parentheses. PPVT, Peabody Picture Vocabulary Test.
† $p < .10$.
* $p < .05$.
** $p < .01$.

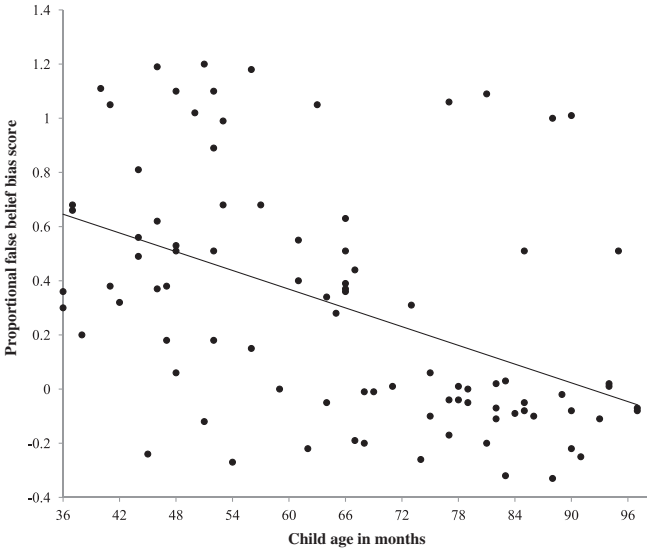

Fig. 3. Mean false belief bias scores on the Sandbox task and children's age in months in Study 2. $R^2 = .14$. Note that false belief bias scores of 1 reflect biased responding, whereas scores of 0 reflect perfect accuracy and non-egocentric responding.

many of the correlations between standard false belief performance and inhibition remained significant.

Discussion

Consistent with Study 1, the results of Study 2 revealed negative correlations between children's age and false belief bias on the Sandbox task. In addition, we obtained negative correlations between Change of Location and false belief trials of the Sandbox task after controlling for memory on the Sandbox task as well as age and receptive vocabulary. These results show that the paper-and-pencil version of the Sandbox task captures meaningful age-related increases in false belief understanding. Further, the Sandbox task conceptually captures the same false belief understanding tapped by a standard measure that shares a similar structure (demonstrating convergent validity). There was no relation between second-order false belief task performance and the false belief trials of the Sandbox task or other standard false belief tasks, demonstrating discriminant validity. This is likely because although the Sandbox task is complex, it still measures only first-order false belief, which is conceptually different from understanding a second-order false belief. Alternatively, the second-order false belief task might have required more vocabulary, memory, and executive processes than the Sandbox task. Importantly, the other first-order false belief tasks also did not correlate with second-order false belief task performance. No relation was found between false belief trials of the Sandbox task and our two measures of inhibition once we controlled for age and vocabulary. Importantly, in all our analyses, we were able to control for memory bias on the Sandbox task, ruling out the possibility that our relations were due to task or story understanding. Rather, we were able to conclude that these relations were driven by false belief misunderstanding specifically.

General discussion

The results of the current studies are encouraging on three fronts. First, young children can complete the paper-and-pencil version of the Sandbox task, and their performance is sensitive to age-related changes. Second, performance on the false belief trials of the Sandbox task relates to false belief performance on a standard false belief task that shares a similar structure (Change of Location task). Third, Sandbox task false belief bias failed to correlate with measures of inhibition. Thus, it is our hope that the Sandbox task will be included in future studies of ToM to detect subtle age-related changes in false belief understanding across childhood.

The paper-and-pencil version of the Sandbox task detected age-related changes in ToM between 3 and 7 years of age. This finding extends previous work documenting age-related changes between 3 and 5 years on the real-object version of the Sandbox task (Sommerville et al., 2013) and that the paper-and-pencil version of the Sandbox task detects false belief bias in adults (Coburn et al., 2015). Moreover, our study showed that the pencil-and-paper version is appropriate for use with young children and can detect age differences, in contrast with its failure to detect age-related increases in a sample of older children (and adults) aged 6–20 years (Begeer et al., 2016). Our findings support previous findings with adults that the Sandbox task has the potential to measure individual differences in ToM ability into later childhood and later adulthood (Bernstein et al., 2011; Sommerville et al., 2013).

Past work has established a relation between the real-object Sandbox task and the standard Change of Location false belief task in preschoolers (Sommerville et al., 2013); however, our study sought to examine the relation between the paper-and-pencil Sandbox task and a wider variety of standard false belief tasks that both do and do not share a similar structure (false belief for an object's location and for an object's identity). Further, we used tasks that asked questions about children's *own* mental states and *others'* mental states to examine relations between thinking about one's own previous states and another's mental state. Performance on the Sandbox task was negatively related to Change of Location performance after controlling for age- and vocabulary-related variance, supporting the idea that the Sandbox task taps into the construct of false belief understanding. This relation suggests convergent validity between the Sandbox and Change of Location tasks because both seem to be mea-

suring a similar construct (false belief understanding as it relates to objects changing location) even though they share a different response format (i.e., binary vs. continuous responses). In contrast, false belief bias on the Sandbox task failed to correlate with the standard false belief tasks that did not share a similar structure (Unexpected Contents and Appearance/Reality) after we controlled for age, vocabulary, and task-specific memory on the Sandbox task. These patterns of correlations suggest discriminant validity given that thinking about one's own or another person's previous belief about an item's contents or identity seems to be distinct from reasoning about another's false belief about an object's previous location. However, there are other differences between the tasks that might account for the lack of relation other than differences between self and other reasoning. One such difference is temporal structure; in the Unexpected Contents and Appearance/Reality tasks, children's questions about the self are always in the past (i.e., "What did you think was inside the box when you first saw it?") versus other questions that are always in the present (i.e., "What would another child think was inside the box if they saw it now?"). Conversely, the Sandbox task poses questions to children about the present only (i.e., "Where will the character look for the item?"). In sum, it seems that false belief bias on the Sandbox task is conceptually similar to the Change of Location task but not to other false belief measures that tap into distinct constructs and have a different structure. These findings suggest that method variance might have contributed to the lack of relations between the Sandbox task and the Unexpected Contents and Appearance/Reality tasks. Future research should further explore the effect of conceptual and methodological differences on correlations among false belief tasks.

We included a second-order false belief task in Study 2 to capture more complex false belief understanding in children over 5 years of age and to avoid ceiling levels of performance on the standard false belief tasks. Nevertheless, we found no relation between performance on the false belief trials of the Sandbox task and second-order false belief performance. Although these two tasks were structurally similar (higher verbal demand and children needed to follow a more complex story with several steps), unrelated performance might be the result of conceptual differences between the Sandbox task and the second-order false belief task. The Sandbox task is a first-order false belief task that requires the representation of another person's mental state, whereas the second-order false belief task requires the representation of a person's mental state of yet another person's mental state. Thus, the lack of relation between these two more complex tasks may support a conceptual distinction between first- and second-order false belief understanding. This bolsters the argument that the Sandbox task has sufficient discriminant validity because performance on this task is distinguishable from tasks that measure a different type of false belief understanding.

Given the well-established link between a wide variety of false belief tasks and inhibitory control (e.g., Carlson, Claxton, & Moses, 2015; Carlson & Moses, 2001; Carlson, Moses, & Breton, 2002; Perner & Lang, 1999; van der Meer, Groenewold, Nolen, Pijnenborg, & Aleman, 2011), an important question from both methodological and theoretical perspectives was whether Sandbox task performance is related to children's inhibition. There was no relation between Sandbox task performance and two well-validated measures of inhibition (Carlson, 2005; Carlson & Moses, 2001; Ponitz et al., 2008; Ponitz, McClelland, Matthews, & Morrison, 2009) in Study 1 and Study 2 after controlling for age and vocabulary. Importantly, the standard false belief measures positively correlated with inhibition, replicating past findings (Carlson & Moses, 2001), and many of these correlations persisted after controlling for children's age in months. Thus, it seems that, compared with standard measures of false belief, the Sandbox task might represent a task that places fewer demands on young children's inhibitory control. Rather than needing to choose the correct option and inhibit the incorrect option as in the standard measures of false belief, the Sandbox task offers a continuous scale on which to respond. This continuous response format might in fact reduce the inhibitory demands that occur at the response level and may be more suitable for measuring false belief understanding in populations who typically struggle with inhibition (e.g., individuals with autism spectrum disorder or attentional problems).

Taken together, these two studies provide important data about the utility of the paper-and-pencil version of the Sandbox task for use with very young children (see Coburn et al., 2015, for its utility with adults). False belief bias scores on the Sandbox task are related to a structurally and conceptually similar measure of standard false belief (but not to structurally and conceptually dissimilar false belief tasks) after controlling for task-specific memory performance. These findings suggest convergent and

discriminant validity of the Sandbox task. False belief bias scores on the Sandbox task were not related to standard measures of inhibition after controlling for age, suggesting that the Sandbox task poses minimal inhibitory demands. Further, the Sandbox task can be used with preschoolers as well as older children without observing floor or ceiling effects. Given that the Sandbox task can include several false belief trials, this task might capture false belief understanding more reliably than single-trial false belief tasks. In future work, researchers should investigate other psychometric properties of the Sandbox task such as reliability.

Acknowledgments

The authors wish to thank Michele Anderson, Prachi Bhuptani, Leia Kopp, Jack Rossing, and Madeline Weissman for assistance with data collection as well as the families and children who participated in the study. Preparation of the manuscript was supported by grants from the Natural Sciences and Engineering Research Council of Canada to C.E.V.M. and C.M.A. and from the Canada Research Chairs Program (950-228407) and the Social Sciences and Humanities Research Council of Canada (435-2013-0291) to D.M.B.

References

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind?". *Cognition, 21*, 37–46.
Begeer, S., Bernstein, D. M., Aßfalg, A., Azdad, H., Glasbergen, T., Wierda, M., & Koot, H. M. (2016). Equal egocentric bias in school-age children with and without autism spectrum disorder. *Journal of Experimental Child Psychology, 144*, 15–26.
Begeer, S., Bernstein, D. M., van Wijhe, J., Scheeren, A. M., & Koot, H. M. (2012). A continuous false belief task reveals egocentric biases in children and adolescents with autism spectrum disorders. *Autism, 16*, 357–366.
Bernstein, D. M., Thornton, W. L., & Sommerville, J. A. (2011). Theory of mind through the ages: Older and middle-aged adults exhibit more errors than do younger adults on a continuous false belief task. *Experimental Aging Research, 37*, 481–502.
Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science, 18*, 382–386.
Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition, 77*, B25–B31.
Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology, 28*, 595–616.
Carlson, S. M., Claxton, L. J., & Moses, L. J. (2015). The relation between executive function and theory of mind is more than skin deep. *Journal of Cognition and Development, 16*, 186–197.
Carlson, S. M., & Meltzoff, A. N. (2008). Bilingual experience and executive functioning in young children. *Developmental Science, 11*, 282–298.
Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72*, 1032–1053.
Carlson, S. M., Moses, L. J., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development, 11*, 73–92.
Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen- through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development, 21*, 315–330.
Coburn, P. I., Bernstein, D. M., & Begeer, S. (2015). A new paper and pencil task reveals adult false belief reasoning bias. *Psychological Research Psychologische Forschung, 79*, 739–749.
Devine, R. T., & Hughes, C. (2013). Silent films and strange stories: Theory of mind, gender, and social experiences in middle childhood. *Child Development, 84*, 989–1003.
Dumontheil, I., Apperly, I. A., & Blakemore, S. J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science, 13*, 331–338.
Dunn, L. M., & Dunn, D. M. (2007). *Peabody picture vocabulary test* (4th ed.). Minneapolis, MN: Pearson.
Flavell, J. H., Flavell, E. R., & Green, F. L. (1983). Development of the appearance–reality distinction. *Cognitive Psychology, 15*, 95–120.
German, T. P., & Leslie, A. M. (2000). Attending to and learning about mental states. In P. Mitchell & K. Riggs (Eds.), *Children's reasoning and the mind* (pp. 229–252). Hove, UK: Psychology Press.
Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance–reality distinction. *Child Development, 59*, 26–37.
Johnson, S., Slaughter, V., & Carey, S. (1998). Whose gaze will infants follow? The elicitation of gaze following in 12-month-olds. *Developmental Science, 1*, 233–238.
Lagattuta, K., Sayfan, L., & Harvey, C. (2013). Beliefs about thought probability: Evidence for persistent errors in mindreading and links to executive control. *Child Development, 85*, 659–674.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.
Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development, 78*, 622–646.
O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development, 67*, 659–677.
Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in Cognitive Sciences, 3*, 337–344.

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that . . .": Attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology, 39*, 437–471.
Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Development, 83*, 469–485.
Ponitz, C. C., McClelland, M. M., Jewkes, A. M., Connor, C. M., Farris, C. L., & Morrison, F. J. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Childhood Research Quarterly, 23*, 141–158.
Ponitz, C. C., McClelland, M. M., Matthews, J. S., & Morrison, F. J. (2009). A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes. *Developmental Psychology, 45*, 605–619.
Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*, 515–526.
Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development, 13*, 73–91.
Roth, D., & Leslie, A. M. (1998). Solving belief problems: Towards a task analysis. *Cognition, 66*, 1–31.
Sommerville, J. A., Bernstein, D. M., & Meltzoff, A. N. (2013). Measuring beliefs in centimeters: Private knowledge biases preschoolers' and adults' representation of others' beliefs. *Child Development, 84*, 1846–1854.
Tahiroglu, D., Moses, L. J., Carlson, S. M., Mahy, C. E. V., Olofson, E. L., & Sabbagh, M. A. (2014). The Children's Social Understanding Scale: Construction and validation of a parent-report theory-of-mind measure. *Developmental Psychology, 50*, 2485–2497.
van der Meer, L., Groenewold, N. A., Nolen, W. A., Pijnenborg, M., & Aleman, A. (2011). Inhibit yourself and understand the other: Neural basis of distinct processes underlying theory of mind. *NeuroImage, 56*, 2364–2374.
Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.
Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*, 655–684.
Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*, 523–541.
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.