Developmental Psychology
2014, Vol. 50, No. 11, 2485–2497

0012-1649/14/$12.00 http://dx.doi.org/10.1037/a0037914

The Children's Social Understanding Scale: Construction and Validation of a Parent-Report Measure for Assessing Individual Differences in Children's Theories of Mind

Deniz Tahiroglu
Ozyegin University and University of Oregon

Louis J. Moses
University of Oregon

Stephanie M. Carlson
University of Minnesota

Caitlin E. V. Mahy
Brock University

Eric L. Olofson
Wabash College

Mark A. Sabbagh
Queen's University

Children's theory of mind (ToM) is typically measured with laboratory assessments of performance. Although these measures have generated a wealth of informative data concerning developmental progressions in ToM, they may be less useful as the sole source of information about individual differences in ToM and their relation to other facets of development. In the current research, we aimed to expand the repertoire of methods available for measuring ToM by developing and validating a parent-report ToM measure: the Children's Social Understanding Scale (CSUS). We present 3 studies assessing the psychometric properties of the CSUS. Study 1 describes item analysis, internal consistency, test–retest reliability, and relation of the scale to children's performance on laboratory ToM tasks. Study 2 presents cross-validation data for the scale in a different sample of preschool children with a different set of ToM tasks. Study 3 presents further validation data for the scale with a slightly older age group and a more advanced ToM task, while controlling for several other relevant cognitive abilities. The findings indicate that the CSUS is a reliable and valid measure of individual differences in children's ToM that may be of great value as a complement to standard ToM tasks in many different research contexts.

Keywords: theory of mind, parent report, measurement, preschool children

Supplemental materials: http://dx.doi.org/10.1037/a0037914.supp

Children's theory of mind (ToM) is typically assessed with a range of laboratory paradigms designed to measure understanding of mental states such as beliefs, desires, emotions, and intentions. Over the last several decades these paradigms have produced a wealth of data concerning developmental changes in mental state understanding and the processes that may underlie them. We now know that striking changes occur during the preschool years in children's appreciation of mental states (Harris, 2006). Throughout this period, ToM develops in a stable, predictable sequence (Wellman & Liu, 2004). By 2 years of age, children's ToM includes a basic understanding of emotion, intention, desire, and perception (Wellman, 2002). Children of this age, however, reveal very little explicit understanding of knowledge and belief. They have difficulty appreciating that people can differ in their beliefs and knowledge states and that someone could believe in something that is false (Wimmer & Perner, 1983), although it remains possible that even infants may at least implicitly recognize something about beliefs (see Baillargeon, Scott, & He, 2010). In addition, they have

This article was published Online First September 29, 2014.

Deniz Tahiroglu, Department of Psychology, Ozyegin University, and Department of Psychology, University of Oregon; Louis J. Moses, Department of Psychology, University of Oregon; Stephanie M. Carlson, Institute of Child Development, University of Minnesota; Caitlin E. V. Mahy, Department of Psychology, Brock University; Eric L. Olofson, Department of Psychology, Wabash College; Mark A. Sabbagh, Department of Psychology, Queen's University, Kingston, Ontario, Canada.

Portions of the research reported here were presented at the 19th annual convention of Association for Psychological Science in 2007, fifth biennial meeting of the Cognitive Development Society in 2007, and biennial meeting of the Society for Research in Child Development in 2009. We would like to thank Mary Rothbart, Mike Posner, Meredith Meyer, Tasha Oswald, and Mary Ann Winter-Messiers for their support in data collection and Adelle Pushparatnam and Eleanor Dizon for their help with data collection and coding.

Correspondence concerning this article should be addressed to Deniz Tahiroglu, Department of Psychology, Ozyegin University, Nisantepe Mah., Cekmekoy, Istanbul 34794, Turkey. E-mail: deniz.tahiroglu@ozyegin.edu.tr

difficulty realizing that appearances may differ from reality (Flavell, Flavell, & Green, 1983) and that people can have different visual perspectives on the same scene or event (Flavell, Everett, Croft, & Flavell, 1981). By the time they are 4 or 5, however, children have a more adult-like understanding of these matters (Harris, 2006).

Laboratory measures of ToM have also revealed a good deal about some of the factors that may affect the development of ToM such as executive functions (Carlson & Moses, 2001), pretend play (Taylor & Carlson, 1997), and language (Milligan, Astington, & Dack, 2007). There may be some purposes, however, for which these tasks are less well suited, at least when used in isolation. In particular, a rather different vein of research assesses the extent to which individual differences in ToM predict concurrent or later facets of development such as social competence, peer relations, and academic achievement (Astington & Pelletier, 1996; Bosacki & Astington, 1999; Caputi, Lecce, Pagnin, & Banarjee, 2012; Dunn & Cutting, 1999; Liddle & Nettle, 2006). In research of this kind, it may be unwise to rely solely on experimental paradigms because they focus on a single informant (the child) tested in a single context (the laboratory) and sometimes with only a single task type (e.g., the false-belief task). A sounder approach is to make use of multiple informants, contexts, and measures. It is well known that multi-informant assessments have the potential to decrease source and setting errors (Merrell, 1999) and that a multitrait–multimethod approach is important in assessing convergent validity (Campbell & Fiske, 1959).

The purpose of the current research was to assess the value of one such additional informant type, namely, parents. We aimed to expand the repertoire of assessment tools available to researchers by developing a reliable and valid parent-report measure of children's ToM: the Children's Social Understanding Scale (CSUS). To be clear, our goal was not to develop a scale that would in any sense replace traditional lab measures of ToM but rather to develop a tool that would complement those measures by providing additional valuable information about children's mental state understanding. In what follows, we first discuss more fully the limitations of current ToM measures for assessing individual differences and then describe preliminary research relevant to the development of the CSUS. Finally, we report three studies assessing the psychometric properties of this new parent-report measure of ToM.

Limitations of Current ToM Measures

Much of the early research on children's mental state understanding focused heavily on paradigms like the false belief task (Cutting & Dunn, 1999; Gopnik & Astington, 1988; Hughes, 1998; Perner, Leekam, & Wimmer, 1987; Wimmer & Perner, 1983). For several reasons, however, relying solely on tasks like this as markers of children's social cognitive skills may be problematic, at least in an individual differences context. First, false-belief measures are pass/fail tasks allowing little room for variance, and very often, children receive only a small number of trials in ToM assessment (Gopnik & Astington, 1988; Moses & Flavell, 1990; Peterson & Slaughter, 2003; Wimmer & Perner, 1983). Given that, it is perhaps not surprising that the reliability of the tasks has varied greatly within and across studies. For example, Hughes et al. (2000) assessed reliabilities for standard and advanced false-belief tasks and found that although internal consistency for *aggregate* task scores was high and test–retest correlations moderate (αs = .84 and .85 for internal consistency in Studies 1 and 2; and *r*s = .77 and .66, for test–retest correlations for standard and advanced task aggregate scores, respectively), test–retest kappa values for the nine individual ToM test questions ranged more widely from poor to moderate levels (.29 to .72, with a mean of .51). Mayes, Klin, Tercyak, Cicchetti, and Cohen (1996) found similarly variable test–retest reliabilities for false-belief test questions (κs ranging from .12 to 1.00, with a mean of .48). The pattern of inconsistent reliability is also evident in samples of children with autism (Grant, Grayson, & Boucher, 2001) and learning disabilities (Charman & Campbell, 1997).

Second, false-belief tasks provide only a limited snapshot of the highly complex and multifaceted nature of mental state understanding. ToM is, of course, much broader than just belief understanding, including, for example, understanding of desire, intention, emotion, perception, and knowledge. This limitation is widely recognized, and many studies now address the problem by assessing or aggregating scores across a variety of mental state tasks (Carlson & Moses, 2001; Hughes & Ensor, 2007; Wellman & Liu, 2004). However, although studies taking this approach have better content coverage than those using single task paradigms, there are nonetheless clear limits to what can be assessed in one or two short sessions with young children. In addition, it is not always feasible to administer a large battery of ToM tasks, especially in research assessing not only ToM but also other developmental capacities.

Third, even when larger batteries are used, ecological validity may not be high. That is, although there is certainly some evidence for the ecological validity of standard ToM tasks (Lalonde & Chandler, 1995; Newton, Reddy, & Bull, 2000; Peskin & Ardino, 2003; Ronald, Happé, Hughes, & Plomin, 2005), they may miss aspects of ToM that could be revealed in children's everyday behaviors, communication, and social interactions (Raver & Leadbeater, 1993). In this regard, naturalistic assessment methods might be used to achieve greater ecological validity (O'Neill, 2007). For example, one approach has been to study the use of mental state terms in children's natural language (e.g., Bartsch & Wellman, 1995; Sabbagh & Callanan, 1998). However, although these studies have certainly provided highly informative data, it is not always feasible to collect large samples of language data in individual differences studies.

Hence, although laboratory ToM tasks are well suited to uncovering the developmental emergence of mental state concepts, it is less clear that by themselves they meet the criteria for good individual differences measures. For that purpose, it is optimal to have multiple measures of the relevant constructs that evidence high reliability and yield considerable variability across participants.

Clearly, additional methods of assessment would be helpful in examining the relation between ToM and other facets of development. In our research, we aimed to expand the measurement tools available to researchers by looking to parents as informants on their children's ToM. A parent-report measure of ToM would potentially allow researchers to assess an unusually wide variety of ToM constructs and might provide greater ecological validity, as parents have the chance to observe their children in many different situations over long periods of time. Although parent reports

certainly have their own limitations (Miller, 1986), they have often been found to be reliable and valid in assessing children's development in other areas, such as temperament (Rothbart, Ahadi, Hershey, & Fisher, 2001), language development (Fenson et al., 1993; O'Neill, 2007), and communication and symbolic behavior (Wetherby & Prizant, 1993). We conjectured that if items were framed appropriately, parents might be able to provide similarly informative assessments of children's mental state understanding.

Preliminary Research

Individual differences in children's ToM may arise for at least two reasons. First, children may differ in the rate at which they achieve various mental state understandings, with some children reaching these milestones early and others late. Second, even after basic concepts have been acquired, children may nonetheless differ in both their propensity to think about mental states and the skill with which they engage in mental state reasoning, and these differences may well persist throughout development. Because these two sources of individual differences are likely to be heavily confounded early in development, and because we wanted a very broad assessment of ToM, the CSUS included items assessing both types.

In preliminary work, a large item pool was generated measuring six core facets of mental state understanding: belief (e.g., understanding that people might have different beliefs about the same situation, that beliefs can be false, and that beliefs can change over time), knowledge (e.g., understanding that people have different levels of knowledge, that knowledge can come from various sources, and that there are levels of certainty in knowledge), perception (e.g., understanding that one can direct others' perceptual attention, that perceptual appearances and reality might not match, and that people might differ in their perceptual access to information), desire (e.g., understanding that people might have different desires, that desires can change over time, and that desires might not always be fulfilled), intention (e.g., understanding that people act based on their intentions, that intentions and outcomes might not match, and that the same intention may result in different outcomes), and emotion (e.g., understanding that people might have different feelings about the same situation, that someone may feel multiple emotions about the same situation, and that facial and vocal expressions reveal emotions). Many of the items were designed to mirror the content of ToM tasks commonly given to children (Flavell et al., 1983; Harris, Donnelly, Guz, & Pitt-Watson, 1986; Moore et al., 1995; Peskin & Ardino, 2003; Schult, 2002; Wellman & Liu, 2004).

This initial pool included items in which the parent was directly asked whether the child possessed a specific aspect of ToM or used relevant mental state terms such as "think," "know," and "feel," (e.g., *Realizes that experts are more knowledgeable than others in their specialty*; *Talks about differences in what people like or want* [e.g., "You like coffee, but I like juice."]), as well as items asking about children's behaviors that might reflect understanding of mental states (e.g., *Tells lies that are really easy to discover*; *Is good at playing hide and seek*; see Appendix for all items in the ultimately developed version of the CSUS).

A small sample of parents was initially interviewed to eliminate items that were too difficult to understand, and the items were then sent to 15 international ToM experts for further evaluation. After additional refinements, a scale consisting of 75 items was constructed. The Children's Social Understanding Scale (CSUS) was then given to a sample of 277 parents (98% mothers) of children between the ages of 2 and 6 at three different data collection sites in North America (Eugene, OR; Seattle, WA; and Kingston, ON, Canada). Parents were asked to rate their children on each item, using a 4-point Likert scale ranging from 1 (*definitely untrue of my child*) to 4 (*definitely true of my child*). Parents were also provided with a "don't know" response option in case they did not have insight into the understanding or behavior tapped in a specific item. Based on item analyses, some items were then discarded and some were revised. In addition, new items were added to expand content coverage, yielding an 80-item scale for the next phase of scale development.

In the current report, we present three studies on further development and assessment of the psychometric properties of the CSUS. We aimed to develop a full version of the scale as well as a short form that might be valuable for researchers simply needing a quick overall assessment of individual differences in ToM in addition to lab assessments. Study 1 reports findings on the construction of the final scales as well as internal consistency, test–retest reliability, and the relation of the CSUS to preschool children's performance on standard lab ToM tasks. In Study 2, data from children and parents from a new sample with a somewhat different set of ToM tasks were collected and analyzed to determine whether the results of the first study would cross-validate. In Study 3, further cross-validation was undertaken and the relation between children's ToM performance and the CSUS was assessed more stringently with several other cognitive abilities controlled.

Study 1

Method

Participants. The 80-item preliminary version of the CSUS was given to 503 parents at three different sites in North America (Eugene, OR; Minneapolis, MN; and Seattle, WA). Reports from 38 parents were excluded because of substantial missing data (more than 20%), leaving data from 465 parents (93% mothers) for final analysis. Participants were recruited by telephoning parents of children between the ages of 2 and 7 included in databases available to researchers at their institutions. The sample was predominantly White and middle-class, reflecting the demographics of the communities from which it was drawn. It consisted of parents of 239 girls (51%) and 226 boys. Children's ages ranged from 28 to 84 months (M_{age} = 50.20, SD = 11.70).

Further data were collected from two subsamples. First, a subset of parents from the main site (in Eugene) was asked to bring children to the lab to assess their performance on standard ToM tasks. Data from two children were excluded because they were unable to complete the tasks. The final sample included 81 preschoolers (40 girls, 41 boys), ranging in age from 37 to 72 months (M_{age} = 53.95, SD = 8.79) and their parents. Second, to assess test–retest reliability, we asked 31 of the parents who attended the lab session also to complete the CSUS 1–4 weeks prior to their lab visit. This final subset consisted of parents of 17 girls and 14 boys. Children's ages ranged from 40 to 66 months (M_{age} = 54.00, SD = 8.30) at the first parental assessment and from 41 to 67 months (M_{age} = 54.45, SD = 8.41) at the second assessment, at

which time children were also tested on the behavioral tasks. To avoid the possibility of practice effects, we used data from the first assessment only in item analyses of the CSUS for the full sample and in analyses of relations to children's task performance.

In sum, we analyzed data from three hierarchically related parent samples: (a) the CSUS responses of the full sample of parents, (b) the CSUS responses of just those parents whose children completed the behavioral assessments of ToM, and (c) the responses of a subset of parents from (b) who completed the CSUS at two different time points.

Measures and procedure. The 80-item version of the CSUS (26 reverse-scored items) consisted of approximately equal numbers of items in each of six subscales (i.e., belief, knowledge, perception, desire, intention, and emotion). Items were quasi-randomly ordered. As in the initial version of the scale, parents were asked to rate their children on a 4-point Likert scale and were again provided with a "don't know" response option. We constructed both boy and girl versions of the scale that varied only in that items with pronouns referring to the child always matched the child's own gender. Completion of the CSUS took about 20 min.

Parents involved only in the scale-construction aspect of the study were mailed questionnaires (71% return rate). Those parents who brought children to the lab for behavioral testing filled out the CSUS during the lab session. Parents in the test–retest sample also completed a mailed CSUS prior to their lab visit.

Children were given three ToM tasks in a single, videotaped session. These tasks were chosen both to maximize variability in the age group and because they reflect fundamental aspects of ToM (i.e., false belief, perception as a source of knowledge, and Level 2 perspective taking). In the contents false-belief task (adapted from Wellman & Liu, 2004), children were shown a Band-aid box that actually contained a toy blue bird. After the children had seen the true content, the box was closed, and children were asked what was really in the box. They were then introduced to a toy figure of a boy who had never seen inside the box. Children were then asked what the boy thought was in the box and whether he had seen inside the box. Children needed to answer both of these questions in addition to the reality question correctly to pass the task. Trials on which children failed the control and/or the reality questions on this and other behavioral tasks were excluded.

In the knowledge-access task (adapted from Wellman & Liu, 2004), children were shown a closed drawer and asked what they thought was inside. The content was then revealed (a toy horse), and the drawer was closed again. Children were next introduced to a toy figure of a girl who had never seen inside the drawer. They were asked whether the girl knew what was in the drawer and whether she had seen inside it.

In the Level 2 perspective-taking task (adapted from Flavell et al., 1981), the experimenter placed a picture of a turtle horizontally on the table such that children saw the turtle right side up, and the experimenter saw it upside down. They were asked whether they themselves saw the turtle right side up or upside down and whether the experimenter saw it right side up or upside down. Children needed to answer both questions correctly to pass the task. As is standard practice in individual differences research, the tasks were given in a fixed order: contents false belief, knowledge access, and Level 2 perspective taking.

Results and Discussion

The CSUS was constructed using the following criteria. First, items with more than 20% missing data were discarded ($n = 3$). Remaining missing values in the scale data (as well as the behavioral data) were replaced using maximum likelihood imputation procedures. Missing data in all of the studies almost always took the form of "don't know" responses. Second, items with low corrected item-total correlations with their subscales ($< .20$) were deleted. Finally, appropriate content coverage and correlations with behavioral ToM tasks were taken into consideration in selecting items.

Following item analysis, we created a 42-item full scale (six reverse-scored items) with seven items in each subscale, as well as an 18-item short form of the CSUS, with three items from each subscale (see Appendix). The 18 items were chosen from the 42-item full scale based on content coverage as well as the items' correlations with behavioral tasks. Both the full and short scales showed excellent internal consistency (αs of .94 and .89, average corrected item-total correlations of .51 and .53, respectively). In addition, test–retest reliability was excellent for both the full and short scales, rs (29) = .88 and .88, ps $< .001$. Note that internal consistency and test–retest reliability were computed after missing data had been replaced with imputed values. The pattern of findings, however, was at least as strong with missing data excluded.

Mean full and short scale scores computed by averaging over items were 3.08 ($SD = 0.45$) and 2.95 ($SD = 0.52$), respectively. Children's age correlated significantly with parent-rated ToM for both the full and short scales, rs (463) = .47 and .50, ps $< .001$, two-tailed, respectively. Correlations among the subscales were high; rs (463) ranged from .55 to .76, ps $< .001$ (average $r = .68$), and remained strong even with age controlled, rs (462) $> .45$, ps $< .001$. As with behavioral ToM tasks (Carlson & Moses, 2001; Hughes & Ensor, 2007; Taylor & Carlson, 1997), these correlations suggest that the subscales cohere as a unified ToM construct. Indeed, an exploratory factor analysis yielded only a large single factor accounting for 32% of the variance, and a subsequent confirmatory analysis revealed that adding subscale-based factors to the model did not appreciably improve fit. Similar patterns were found in our subsequent studies, and hence, for the most part, we do not proceed further with subscale analyses (subscale means, reliabilities, and correlations with task performance for all of the studies may be found in Tables S1 to S5 of the online supplemental materials).

Table 1 shows means and standard deviations for the three behavioral ToM tasks as well as a ToM composite formed by averaging scores across the three tasks. Consistent with the literature (Harris, 2006), children's performance on the individual and composite behavioral tasks correlated significantly with age, rs (79) $> .44$, ps $< .001$. Also in line with the literature (e.g., Wellman & Liu, 2004), children performed significantly better on the knowledge access task than on the false-belief and Level 2 perspective-taking tasks, ts (80) > 4.08, ps $< .001$ (Cohen's ds $> .45$). Children's performance on the false-belief and Level 2 perspective-taking tasks did not differ. False-belief performance correlated significantly with performance on both the knowledge access, $r(79) = .57$, $p < .001$, and perspective-taking tasks, $r(79) = .25$, $p = .023$. Scores on the knowledge access and perspective-taking tasks were not significantly related.

Table 1
Means (and Standard Deviations) of Tasks in Studies 1, 2, and 3

Task	Study 1 Scale construction sample ($N = 81$)	Study 2 Validation sample ($N = 94$)	Study 3 Validation sample ($N = 123$)
Knowledge access	.75 (.43)		
Contents false belief	.48 (.48)	.34 (.44)	
Level 2 PT (turtle)	.48 (.50)	.54 (.49)	
Explicit false belief		.50 (.43)	
Level 2 PT (book)		.64 (.46)	
AR (Rock/sponge)		.48 (.48)	
AR (Red/black castle)		.61 (.47)	
ToM composite	.57 (.35)	.52 (.28)	
Restricted view[a]			7.77 (2.04)
Prospective memory			3.00 (1.50)
Backwards digit span			1.60 (1.50)
Truck loading			2.72 (1.42)

Note. Scores ranged from 0 to 1 for tasks in Studies 1 and 2. Range of possible scores was 0 to 10 for the restricted-view ToM (theory of mind) task, 0 to 4 for prospective memory task, 0 upwards for backwards digit span task, and 0 to 4 for truck loading task. PT = perspective taking; AR = appearance–reality.
[a] $N = 92$ for the restricted-view task.

Most importantly, children's performance on the behavioral ToM tasks correlated with parent reports of their ToM. Table 2 shows raw and partial correlations controlling for age between the CSUS and the individual and composite ToM variables. The ToM composite lab score correlated significantly with both the full and short parent-report scales, rs (79) = .31 and .37, ps = .005 and .001, respectively, as did both the false-belief and knowledge-access measures. In contrast, performance on the Level 2 perspective-taking task did not correlate significantly with either the full or short scales. Finally, when age was held constant, the correlations between the behavioral ToM measures and the parent-report scales fell below significance, although most remained in the predicted direction.

Summary

We were able to construct psychometrically sound full and short versions of the CSUS based on parent-reported ToM from 465 parents. The scales showed very high internal consistency and test–retest reliability. In addition, they predicted children's performance on behavioral tasks, thus providing important evidence of validity. That said, the correlations fell below significance when age was held constant, a point to which we return in Studies 2 and 3.

Study 2

Item selection in Study 1 was based in part on maximizing the internal consistency of the scales and their correlations with ToM task performance. The psychometric strength of the scales could thus have been artificially inflated through capitalization on correlations that were high by chance. To obtain an unbiased estimate of the internal consistency and validity of the CSUS, we therefore conducted a cross-validation study with a new sample of parents and children using a somewhat different set of ToM tasks.

Method

Participants. One hundred and two parents were recruited as part of other ongoing studies. Parents with more than 20% missing data ($n = 8$) were excluded, leaving a final sample of 94 parents (85% mothers). Behavioral data from children of these 94 parents were also collected. The child sample ranged in age from 37 to 76 months ($M_{age} = 52.65$, $SD = 9.58$) and included 45 girls and 49

Table 2
Raw and Age-Controlled Scale–Task Correlations in Studies 1, 2, and 3

	Study 1 Scale construction sample ($N = 81$)		Study 2 Validation sample ($N = 94$)		Study 3 Validation sample ($N = 92$)	
Task	Full scale	Short scale	Full scale	Short scale	Full scale	Short scale
Knowledge access	.28* (.13)	.35** (.20†)				
False belief	.26* (.12)	.32** (.17)	.38** (.30**)	.42**(.35**)		
Level 2 PT	.15 (−.03)	.17 (−.03)	.35** (.30**)	.38** (.34**)		
AR			.22* (.15)	.28* (.22*)		
Restricted-view ToM					.43** (.34**)	.43** (.32**)
ToM composite	.31** (.11)	.37** (.17)	.42**(.35**)	.47** (.41**)		

Note. Partial correlations controlling for age are shown in parentheses. PT = perspective taking; AR = appearance–reality task; ToM = theory of mind.
† $p < .10$. * $p < .05$. ** $p < .01$ (2-tailed).

boys. Parent reports and behavioral data were collected at two different sites in North America (Eugene, OR, and Crawfordsville, IN).

Measures and procedure. Parents completed the CSUS either in their homes or during lab visits (the version given to parents was the same as that in Study 1, although analysis was restricted to just the 42 items included in the final scale). Some children completed behavioral tasks in a quiet room in their preschools ($n = 52$), whereas others did so in the lab ($n = 42$). Children were given two false-belief tasks (i.e., contents false-belief and explicit false-belief tasks, adapted from Wellman & Liu, 2004), two Level 2 perspective-taking tasks (i.e., the turtle task, adapted from Flavell et al., 1981, and the book task, adapted from Carlson, Mandell, & Williams, 2004), and two appearance–reality tasks (i.e., the rock–sponge and red castle tasks, adapted from Flavell, Green, & Flavell, 1986). The tasks were given in a fixed order: Contents false-belief task, Level 2 perspective-taking turtle task, appearance–reality (sponge–rock) task, explicit false-belief task, appearance–reality (red–black castle) task, and Level 2 perspective-taking book task.

The contents false-belief task was the same as that in Study 1. In the explicit false-belief task, children were told a story (accompanied by pictures) about a boy who was looking for his mittens that could be located either in a backpack or a closet. They were told that his mittens were really in the backpack but that the boy thought they were in the closet. Children were then asked where he would look for his mittens and where they really were. Children needed to answer both questions correctly to pass.

The Level 2 perspective-taking turtle task was the same as that in Study 1. The Level 2 perspective-taking book task was similar and involved a picture book that was right side up from the child's perspective but upside down from the experimenter's perspective.

For the appearance–reality tasks, children were shown two objects with misleading appearances. The rock–sponge task involved a discrepancy between a real and apparent identity (e.g., a sponge that looked like a rock), and the red castle task involved a discrepancy between real and apparent color (e.g., the red castle looked black when behind a green filter). Children were questioned in the standard way about the object's appearance and its reality. They passed the tasks only if they answered both questions correctly.

Results and Discussion

Internal consistency was again excellent (αs of .90 and .84, average corrected item-total correlations of .41 and .44, for the full and short scales, respectively). Children's age again correlated significantly with parent-reported ToM: full scale, $r(92) = .29$, $p = .005$, and short scale, $r(92) = .26$, $p = .011$.

Table 1 shows means and standard deviations of the behavioral tasks in Study 2. We averaged scores on each of the two false-belief, perspective-taking and appearance–reality tasks and also formed an overall ToM composite by averaging scores on all six tasks. False-belief, perspective-taking, and appearance–reality task aggregate scores were significantly interrelated, *r*s (92) ranging from .25 to .47, *p*s <. 05. In addition, children's age correlated with performance on the three behavioral task aggregate scores, *r*s (92) > .23, *p*s < .05, as well as the overall ToM composite, $r(92) = .39$, $p < .001$.

Table 2 displays correlations between the CSUS and children's performance on the ToM tasks. As in Study 1, the ToM composite lab score correlated significantly with the full and short versions of the scale, *r*s (92) = .42 and .47, *p*s < .001, respectively, as did performance on each of the ToM measures taken separately. Strikingly, the parent report–task correlations were as high, and in some cases higher, than the correlations among the ToM tasks themselves.

Correlations between parents' ratings and behavioral performance were also assessed with age controlled (see Table 2). Critically, parents' ratings on the full and short scales predicted children's ToM performance (as measured by the composite score) over and above age, *r*s (91) = .35 and .41, $p = .001$ and $p < .001$, respectively. Age-controlled correlations between the scales and scores on the individual behavioral measures also remained significant with the exception of the relation between appearance-reality and the full scale.

Summary

In Study 2, the CSUS again showed excellent internal consistency. In addition, moderately high correlations with behavioral ToM task performance were found, providing evidence of scale validity. We also computed correlations controlling for age to examine the possibility that parents' reports were influenced mainly by perceptions of age-related general development rather than specific ToM competence. Unlike in Study 1, parent-reported ToM for the most part significantly predicted children's behavioral task performance over and above age in this sample.

Notably, the scale–task correlations (with and without age controlled) were higher in this sample than they were in Study 1. We had expected correlations to decline somewhat because in Study 1 items were in part selected to maximize relations to behavioral ToM performance. The fact that they did not decline may reflect that the number of tasks was increased in Study 2 (six vs. three test trials), thereby providing a more reliable assessment of behavioral ToM than in Study 1.

Study 3

In Study 3, we aimed to further cross-validate the CSUS. In Study 2, the CSUS was related to children's ToM performance independently of age, providing evidence that parents were not simply assessing their children in terms of some crude maturational metric that also happened to relate to ToM. In Study 1, however, most of the significant correlations with behavioral performance had disappeared with age controlled. The first goal in Study 3 then was to provide an additional assessment of whether relations with task performance would be found independent of age. A second and related goal was to assess whether these relations would be found independent of other important aspects of cognitive functioning. We took advantage of a separate study we were conducting examining relations among prospective memory, executive functioning (working memory and planning), and ToM in 4- and 5-year-olds (Mahy & Moses, in press). All of these cognitive abilities develop considerably in the preschool years and have been found to relate to ToM performance (e.g., Carlson & Moses, 2001; Carlson, Moses, & Claxton, 2004; Ford, Driscoll, Shum, & Macaulay, 2012; Mahy & Moses, 2011). If parents'

assessments of their children's ToM were simply driven by global perceptions of children's cognitive competence, then relations between the CSUS and ToM should disappear with these other cognitive abilities controlled. A third and final goal was to assess whether our earlier findings would generalize to a ToM task with a somewhat different structure than those employed in Studies 1 and 2. Specifically, children completed variants of the restricted-view task, which yield greater variance in 4- and 5-year-olds than do the simpler ToM tasks used in the earlier studies (Chandler & Helm, 1984; Taylor, Cartwright, & Bowden, 1991).

Method

Participants. One hundred and thirty-four parents and their children were recruited at the main data collection site (Eugene, OR). Parents with more than 20% missing data ($n = 11$) were excluded, leaving a sample of 123 parents (86% mothers) and 123 children (age range 46 to 72 months, $M_{age} = 58.54$, $SD = 7.55$; 64 girls and 59 boys). Ninety percent of the children were White, 7% of mixed ethnicity, 2% Hispanic, and 1% of Asian descent.

Measures and procedure. Parents filled out the 42-item CSUS during lab visits while children were completing the behavioral measures. The tasks were given in a fixed order: prospective memory, backward digit span as a measure of working memory, truck-loading task as a planning measure, and the restricted view task as a ToM measure.

Prospective memory task. This task required children to name objects depicted in four stacks of cards and to provide a novel response to certain target cards (adapted from Mahy & Moses, 2011). They were asked to help Morris the Mole, a stuffed animal, learn what was on the cards by naming the pictured objects but to hide any animal cards in the box behind them because Morris was afraid of animals. Cards were 3- × 3-inch color pictures of everyday objects (e.g., food, furniture, toys, animals). Before card sorting, children completed either a scrambled or unscrambled version of the self-ordered pointing task (Hongwanishkul, Happaney, Lee, & Zelazo, 2005) as part of a manipulation for the separate study. Subsequently, children alternated between drawing a picture for 1 min and naming a stack of cards until they had named all four stacks. Each stack contained 12 cards including one animal card. Children were given a score out of 4 based on the number of animal cards they correctly placed in the box.

Backward Digit span. To measure working memory, children were asked to complete the Digits Forward and Digits Backward subscales from the Wechsler Intelligence Scale for Children (3rd ed., WISC–III; Wechsler, 1991). We used the backward digit span task as our measure of working memory but administered the forward digit span to familiarize children with the task. In the digits backward task, children were asked to repeat a series of numbers in backward order. They began with two numbers, and after they completed two trials successfully, an additional number was added. The task ended when they failed two consecutive trials. The backward digit span score was the number of digit strings children repeated accurately.

Truck loading. In this planning task, children delivered party invitations to a neighborhood of houses (Carlson et al., 2004). The color of the invitations corresponded to the color of cardboard houses lining the street. Children were asked to place the invitations into a truck and then deliver them while following three rules: (a) drive the truck in one direction on the street, (b) deliver the invitations as fast as possible to the appropriate color-matched houses, and (c) deliver invitations only from the top of the pile in the back of the truck. To succeed at the task, children thus needed to place the invitations in the truck in the reverse order to which they were to be delivered. The experimenter demonstrated a trial, and children were then given a practice trial with two houses. They were asked to deliver the invitations to houses on the block twice, and if they successfully delivered the invitations on at least one of the trials, an additional house was added. There were four levels of trials ranging from two houses to five houses. If children failed two trials at the same level consecutively, the task ended. Children were scored on their highest level of achievement from 0 (*could not deliver invitations to two houses correctly*) to 4 (*delivered invitations to five houses correctly*).

Restricted view task. In this task (adapted from Taylor et al., 1991), children were asked to guess the contents of several pictures. Each picture was mounted on card stock, and a sheet of blue card stock was taped to one edge to serve as a removable cover. A small rectangular opening was cut in each cover. The extent to which the picture could be seen when the cover was in place varied to create three types of stimuli: (a) identifiable—a sufficient part of the object showed to allow identification of the picture; (b) empty—no part of the object could be seen; that is, the view showed empty white paper, and (c) nondescript/ambiguous—only a small nondescript part could be seen. There were two identifiable stimuli (dog/girl), two empty stimuli (turtle/bunny), and one nondescript stimulus (deer).

After children guessed, the cover was removed, and they were asked, "What is it a picture of?" The cover was then replaced, and they were asked, "If another child about your age came into this room right now, would that child know what this is a picture of?" They were then asked, "At the beginning, before I took the cover off, could you know that there was a _____ in the picture?" The five trials were administered in a fixed order: identifiable, empty, identifiable, empty, and nondescript. In identifiable trials, children should state that another child would know the identity of the picture and that they themselves had known it before the cover was removed. In empty and nondescript trials, they should state that neither they nor the other child could know the identity. Children were given a score from 0 to 10 depending on the number of correct answers they provided in each of the five trials. Because 31 children either guessed correctly at the beginning that there was a deer in the picture on the nondescript trial or did not answer the first question at all, only those who guessed incorrectly ($n = 92$) were included in the main analysis involving children's behavioral performance (the full sample of 123 parents continued to be used in analyses of scale properties). The 31 excluded children did not differ from those who were included on prospective memory ($p = .76$), working memory ($p = .76$), planning ability ($p = .78$), or overall CSUS scores ($ps = .23$ and .28, for full and short scales, respectively).

Results and Discussion

Internal consistencies of the full and short scales remained very high (αs of .91 and .81, average corrected item-total correlations of .43 and .41, respectively). Children's age once again correlated

significantly with the full and short scales, *r*s (121) = .27 and .30, *p*s = .002 and .001, respectively.

Table 1 shows means and standard deviations of the behavioral tasks in Study 3. Prospective memory, working memory, and planning were all significantly intercorrelated, *r*s (111 to 116) > .35, *p*s < .001, and also related to the restricted view ToM task, *r*s (88 to 90) > .37, *p*s < .001. Children's age correlated with prospective memory, working memory, and planning, *r*s (114 to 121) > .28, *p*s < .01, as well as with the restricted-view ToM task, $r(90) = .49$, $p < .001$.

Table 2 displays correlations between the CSUS and children's performance on the behavioral tasks. As in Studies 1 and 2, the composite score on ToM test trials (out of 10 points) correlated significantly with the full and short scales, *r*s (90) = .43 and .43, *p*s < .001. It is important to note that when age was controlled, ToM was still significantly related to the full and short scales, *r*s (89) = .34 and 32, *p*s = .001 and .002, respectively. Because of substantial missing data on the nondescript trial of the restricted view ToM task, we re-analyzed the data using only the identifiable and empty trials. Similar patterns emerged: raw *r*s (115) = .36 and .36 (*p*s < .001), and age-controlled *r*s (114) = .28 and .26 (*p*s = .003 and .004), for the full and short scales, respectively.

To further assess construct validity, we examined relations between the CSUS and the other cognitive tasks. If the CSUS is indeed measuring children's ToM, then it should relate more strongly to children's performance on ToM tasks than to performance on the other behavioral tasks. This was certainly the case for planning and prospective memory but not for working memory. Specifically, the relations between planning and prospective memory and the full-scale CSUS failed to reach significance, *r*s (114 and 121) = .17 and .15, *p*s = .062 and .096, respectively, and were significantly lower than the relation between the restricted-view task and the CSUS, *Z*s > 2.54, *p*s < .01. In contrast, working memory was significantly related to the CSUS, $r(116) = .33$, $p < .001$, although not quite as highly as was the restricted-view ToM task, $r(90) = .43$, $p < .001$. The two correlations did not differ significantly ($Z = 1.05$, $p = .15$). Similar patterns emerged for the analysis in relation to the short scale CSUS.

Importantly, however, when we controlled not only for age but also for working memory, planning, and prospective memory, children's ToM remained significantly related to both the full and short scales, *r*s (83) = .31 and .28, *p*s = .005 and .01, respectively.

Summary

In Study 3, the CSUS again showed strong internal consistency. In addition, with a sample of slightly older children than those in the earlier studies and a relatively more advanced ToM task, parents' CSUS ratings once more correlated with children's ToM performance. It should be noted that, as in Study 2, the relation remained significant with age held constant. Study 3 provided stronger evidence of construct validity, however, by showing that the relation between parent-reported ToM and children's ToM performance remained significant even when several other cognitive abilities were also controlled. The CSUS thus appears to be tapping into specific mental state understanding as opposed to generic cognitive competence.

Combined Results

In a final analysis, we combined the parent data from all three studies to assess some additional general properties of the CSUS. First, we looked for gender effects. Differences between girls and boys are typically not found on ToM tasks but when they are, they tend to favor girls (e.g., Charman, Ruffman, & Clements, 2002; Cutting & Dunn, 1999). Consistent with these behavioral findings, we obtained small but significant gender differences. Girls were rated significantly higher than boys on the full ($M = 3.18$, $SD = 0.41$, vs. $M = 3.09$, $SD = 0.44$) and short scales ($M = 3.07$, $SD = 0.47$, vs. $M = 2.96$, $SD = 0.51$), *t*s(680) = 2.73, and 2.77, *p*s = .006, respectively. Effect sizes were small (Cohen's *d*s of .21 and .22, respectively).

Second, we assessed differences in mean performance on the subscales. Preschool children typically show greater understanding of some mental states, such as desire, than of others, such as belief (Wellman & Liu, 2004). As noted earlier, factor analyses of the CSUS indicated one general ToM factor as opposed to separate subscale factors. Nonetheless, it was possible that differences in parent-reported ToM proficiency might emerge across the subscales. Table 3 shows the relevant means for the combined data. Consistent with prior behavioral findings, when ratings on the subscales were examined, parents reported significantly weaker understanding of belief than understanding of each of the other mental states, *t*s (681) > 4.91, *p*s < .001 (Cohen's *d*s ranging from .20 to .53). They also reported significantly weaker understanding of perception than of knowledge, desire, intention, and emotion, *t*s (681) > 5.06, *p*s < .001 (Cohen's *d*s ranging from .20 to .28). Parent-reported understanding for knowledge, desire, intention, and emotion did not differ.

The items in each scale tapped concepts of varying complexity. Comparing subscale means for each mental state is therefore a crude indication of difficulty in understanding across mental states. However, one item type was very closely matched across scales, specifically the item asking whether children talked about each of the mental states. When these specific items were analyzed (see Table 3 for means and standard deviations), talk about belief was rated by parents as significantly less characteristic of their children than talk about each of the other mental states, *t*s (681) > 4.82, *p*s < .001 (Cohen's *d*s varying between .20 and .54). Talk about

Table 3
Means (and Standard Deviations) of Scales and Mental State Talk Items in Combined Sample

	Combined sample ($N = 682$)	
Variable	Scale	Mental state talk item
Belief	2.97 (0.64)	3.28 (0.84)
Knowledge	3.18 (0.56)	3.42 (0.73)
Perception	3.08 (0.47)	3.63 (0.63)
Desire	3.19 (0.45)	3.74 (0.57)
Intention	3.19 (0.49)	3.44 (0.77)
Emotion	3.20 (0.44)	3.61 (0.64)
Full scale	3.14 (0.43)	
Short scale	3.02 (0.49)	

Note. Mean subscale and scale scores were computed by averaging over items.

knowledge and talk about intention were also rated less characteristic of children than talk about perception, desire, and emotion, *t*s (681) > 5.74, *p*s < .001 (Cohen's *d*s varying between 21 and .44). Talk about emotion and talk about perception were rated less characteristic of children than talk about desire, *t*s (681) > 3.71, *p*s < .001 (Cohen's *d*s = .22 and .15, respectively). Putting these findings together, while parents reported that their children talked the least about belief, they reported that they talked the most about desire. These findings are very much consistent with what has been observed in children's natural language data (e.g., Bartsch & Wellman, 1995; Wellman, Harris, Banerjee, & Sinclair, 1995; Wellman, Phillips, & Rodriguez, 2000).

Finally, three items in the CSUS closely matched the content of tasks that Wellman and Liu (2004) used in their scaling of behavioral ToM performance: diverse desires, explicit false belief, and real–apparent emotion. When items tapping these specific concepts were analyzed, their ordering of difficulty was identical to that found in Wellman and Liu's scaling. Understanding of diverse desires (Desire subscale Item 3 in Appendix) was rated significantly more characteristic of children ($M = 3.55$, $SD = 0.71$) than understanding of explicit false beliefs (Belief subscale Item 3; $M = 3.23$, $SD = 0.87$), which in turn was rated more characteristic of children than understanding of real–apparent emotion (Emotion subscale Item 2; $M = 1.95$, $SD = 0.88$), *t*s (681) > 10.30, *p*s < .001 (Cohen's *d*s .40 and 1.46, respectively).

General Discussion

The goal of this research was to construct a reliable and valid parent-report measure of ToM as an additional tool for assessing social understanding in children. As reported in three studies, data from several different samples revealed strong evidence of reliability and validity for the CSUS.

In Study 1, we constructed a 42-item full scale and an 18-item short scale, both of which showed high internal consistency as well as strong test–retest reliability across 1–4 weeks. These psychometric properties of the CSUS are as high as or higher than those of the behavioral ToM tasks discussed in the introduction (Charman & Campbell, 1997; Grant et al., 2001; Hughes et al., 2000; Mayes et al., 1996).

Study 1 also provided initial evidence with respect to the validity of the CSUS, although this issue was addressed more fully in the later studies. First, parents' ratings correlated with children's age. As children mature, not only do they perform better on behavioral ToM tasks but also parents' ratings of their ToM reveal corresponding improvement. Second, evidence of convergent validity emerged in that children's performance on ToM tasks generally correlated significantly with their parents' ratings, although these correlations fell below significance when age was controlled.

Of course, in Study 1, items were selected in part to jointly maximize internal consistency and relations to ToM (although not test–retest reliability or correlations with age). As is always the case in scale construction, these item analytic procedures may have spuriously inflated some of our estimates of reliability and validity. In Studies 2 and 3, we tested this possibility in two different cross-validation samples, in which these estimates were obtained independently of item selection procedures.

Critically, the psychometric properties of the CSUS were generally maintained and, in some respects, strengthened in Studies 2 and 3. First, both scales again showed high internal consistency. Second, parents' ratings once more significantly correlated with age. Third, and most importantly, parents' ratings again significantly correlated with children's ToM performance. Finally, in contrast to Study 1, these relations remained significant even after controlling for age in both Studies 2 and 3, ruling out the possibility that parents were simply judging their children's ToM in some crude fashion as a function of maturational status.

Interestingly, the correlations between the CSUS and children's ToM performance (with and without age controlled) were generally higher in Studies 2 and 3 than in Study 1. These findings may have been due to the differing numbers of ToM test trials administered in the sessions (three trials in Study 1 vs. six trials in Study 2 and five trials in Study 3). The somewhat more thorough assessment of ToM in the later studies may have generated a stronger pattern of findings.

Study 3 also provided evidence of construct validity for the CSUS. Specifically, we assessed other aspects of children's cognitive functioning that are known to relate to behavioral ToM: working memory, planning, and prospective memory abilities (Carlson et al., 2004; Carlson, Moses, & Breton, 2002; Ford et al., 2012). We replicated these earlier behavioral findings: Children's performance on our ToM tasks correlated significantly with working memory, planning, and prospective memory. The CSUS revealed a somewhat similar pattern of findings, correlating significantly with working memory, although not with planning and prospective memory. Critically, however, the CSUS remained significantly related to children's ToM performance, even with all of these cognitive abilities controlled in addition to age. The CSUS appears then to be more than an index of general cognitive functioning.

It might be argued that despite the statistical significance of our correlations, they are too small to be of importance (full-scale CSUS–ToM composite correlations of .31, .42, and .43 across the three studies). In response, we would note that there are constraints on the likely size of such correlations. First, the CSUS assesses a plethora of mental state understandings, whereas the ToM tasks used in our studies and others assess only a few. As a result, very high concordance between the two assessments would not be expected because they are designed to measure different, albeit overlapping, sets of constructs. Second, as mentioned in the introduction, the reliability of standard ToM tasks is variable, and so to the extent that CSUS–ToM correlations are not perfect, it is unlikely that the blame lies entirely with the parent-report measure. Finally, our correlations are at least as strong as those found among child measures of ToM themselves in our data and in other literature (Carlson & Moses, 2001; Carlson et al., 2002; Cutting & Dunn, 1999; Taylor & Carlson, 1997). In our view, this pattern is an especially strong indicator of the validity of the CSUS. More typically, correlations across different informants are substantially weaker than those obtained within informants of a single type. For example, only modest correlations have been found between parental reports of children's impulsivity and inhibitory control and their performance on behavioral tasks measuring these constructs (Carlson & Moses, 2001; Kochanska, Murray, Jacques, Koenig, & Vandegeest, 1996).

An additional source of validity comes from our analysis of scale characteristics for the complete parent sample. First, girls were rated significantly higher than boys on the full and short

version of the CSUS. As noted earlier, gender differences are typically not found in the behavioral literature but when they are they tend to favor girls (Charman et al., 2002; Cutting & Dunn, 1999). This pattern suggests a relatively small gender difference. Our findings are consistent with this interpretation: With our combined sample of 682 parents, we detected a difference, but the effect size was small (Cohen's *d*s of .21 and .22 for the full and short scales, respectively). Second, and more importantly, consistent with the literature on children's task performance (Wellman & Liu, 2004), parents reported greater understanding for some mental states (e.g., desire) than for others (e.g., belief). The CSUS thus mirrors aspects of the sequence of ToM development commonly observed in behavioral performance.

A rather different point to note is that the relations between parent-reported ToM and children's ToM performance not only demonstrate the validity of the CSUS but also bolster the external validity of the behavioral tasks themselves. To the extent that standard lab assessments reflect children's everyday ToM, some concordance should be found between performance on those assessments and how parents view their children's social cognitive understanding. Our findings provide solid evidence of such concordance and thus fortify the case for the ecological validity of lab ToM assessments (Lalonde & Chandler, 1995; Newton et al., 2000; Peskin & Ardino, 2003; Ronald et al., 2005).

Finally, there are of course limitations to parent reports as a source of information about children's mental state understanding. Parents presumably vary in their ability to infer what their children do and do not understand, their memories of what children say and do are surely fallible, some parents may wish to present their children in a good light and so overestimate their abilities, and others may be seeking help managing children's perceived problems, leading them to exaggerate their children's lack of understanding. For these reasons we would not recommend treating CSUS scores as an *absolute* index of the ages at which specific mental state understandings emerge. Rather, the CSUS is best thought of as a broad measure of children's *relative* standing on ToM, and for that reason, it is best suited to individual-differences research on relations between ToM and other aspects of development. In this respect, the CSUS differs appreciably from behavioral scales of ToM (Wellman & Liu, 2004) and emotion understanding (Pons & Harris, 2005) which are explicitly designed to assess when different conceptual understandings come online in a more absolute sense.

Future Directions

In future research, it will be important to examine other aspects of reliability and validity. For example, we tested short-term test–retest reliability over a period of weeks in Study 1. We do not know how stable scale scores would be over longer time periods. It would also be important to test the 18-item short version of the CSUS on its own to determine whether its psychometric properties hold up outside the context of the full scale.

Relatedly, our samples were mostly White and middle-class. Although we did not collect extensive demographic information from participants in our studies, we are confident that these samples are representative of those parents and children who typically participate in standard ToM research. Nevertheless, conducting this research in economically, educationally, and culturally more diverse populations is needed to assess whether the CSUS is a widely applicable measure.

As further evidence of construct validity, it will be important to assess how the CSUS relates to other aspects of children's development. For example, we know that ToM task performance relates to such constructs as social competence (Bosacki & Astington, 1999; Lalonde & Chandler, 1995), verbal ability (Milligan et al., 2007), executive function (Carlson & Moses, 2001), and general cognitive ability (Carlson et al., 2002). We need to determine whether parent-reported ToM shows similar relations to these constructs (as measured behaviorally or by parent report), and further, whether relations between parent-reported ToM and children's ToM task performance persist when these other factors are held constant. Study 3 provided initial positive evidence on this front with respect to aspects of executive function (working memory and planning), but further work is necessary.

It will also be important to examine how well the CSUS taps ToM deficits in atypical populations such as individuals with autism. If the CSUS is a valid measure of ToM, then it should discriminate between typical and atypical groups. Initial data from our lab suggest that it does. Parents rated typically developing children ($n = 18$; age range = 10–16 years) higher than children with autism spectrum disorder (ASD; $n = 15$; age range = 10–16 years) on the CSUS, even after controlling for effects of intelligence and age. Moreover, scores on the CSUS generally predicted the severity of autistic traits as measured by the Autism Quotient (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001) and the Asperger Syndrome Diagnostic Scale (Myles, Bock, & Simpson, 2001) for both the typically developing and ASD groups, even after age and IQ were controlled. If these findings persist in a larger sample, they will be an important indicator of CSUS validity.

In addition, more fine-grained data at the subscale level are needed. In our studies, the subscales were highly correlated, cohering as a single general dimension in factor analyses. Hence, it is difficult to determine fully the extent to which parents were providing a global assessment of ToM as opposed to highly differentiated information about understanding of each mental state. That said, important evidence for the latter comes from the fact that parents rated their children's understanding of some mental states lower than others in ways roughly commensurate with children's typical ToM performance (Wellman & Liu, 2004). Further evidence in this regard could be obtained by assessing whether parent-reported understanding of one mental state (e.g., belief) relates more strongly to performance on belief tasks than to performance on, for example, desire tasks (and vice versa).

Finally, most of our validation data come from preschoolers in the 3- to 6-year old age range. Thus, we cannot say for certain how the scale would perform beyond this age range. However, preliminary data from the autism study mentioned earlier reveal that even typically developing children at older ages do not score at ceiling. Thus, it is possible that with appropriate modifications to item wordings, the CSUS could be used effectively to measure individual differences in ToM in older children and adolescents as well. In addition, assessing the CSUS in older children would provide an opportunity to disentangle the two forms of individual differences mentioned in the introduction: "maturational" differences in the points at which children develop specific understandings versus "dispositional" differences in the propensity to and skill with

which children make use of those understandings. Although these theoretically different sources of variability are difficult to tease apart early in development, they should begin to diverge as children age.

Conclusion

Our findings indicate that parents can indeed provide useful information about their children's ToM. The CSUS appears to be a reliable and valid measure of ToM, although further demonstrations of its psychometric properties will be helpful. We believe the CSUS will be an invaluable tool for correlational and longitudinal research examining individual differences in children's ToM and its relations to other aspects of development.

References

Astington, J. W., & Pelletier, J. (1996). The language of mind: Its role in learning and teaching. In D. R. Olson & N. Torrance (Eds.), *The handbook of education and human development: New models of learning, teaching and schooling* (pp. 593–619). Oxford, United Kingdom: Blackwell.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14,* 110–118. doi:10.1016/j.tics.2009.12.006

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger syndrome/high functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31,* 5–17. doi:10.1023/A:1005653411471

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind.* New York, NY: Oxford University Press.

Bosacki, S., & Astington, J. W. (1999). Theory of mind in preadolescence: Relation between social understanding and social competence. *Social Development, 8,* 237–255. doi:10.1111/1467-9507.00093

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105. doi:10.1037/h0046016

Caputi, M., Lecce, S., Pagnin, A., & Banerjee, R. (2012). Longitudinal effects of theory of mind on later peer relations: The role of prosocial behavior. *Developmental Psychology, 48,* 257–270. doi:10.1037/a0025402

Carlson, S. M., Mandell, D. J., & Williams, L. (2004). Executive function and theory of mind: Stability and prediction from age 2 to 3. *Developmental Psychology, 40,* 1105–1122. doi:10.1037/0012-1649.40.6.1105

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72,* 1032–1053. doi:10.1111/1467-8624.00333

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. doi:10.1002/icd.298

Carlson, S. M., Moses, L. J., & Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *Journal of Experimental Child Psychology, 87,* 299–319. doi:10.1016/j.jecp.2004.01.002

Chandler, M. J., & Helm, D. (1984). Developmental changes in the contributions of shared experience to social role taking competence. *International Journal of Behavioral Development, 7,* 145–156. doi:10.1177/016502548400700203

Charman, T., & Campbell, A. (1997). Reliability of theory of mind task performance by individuals with a learning disability: A research note. *Journal of Child Psychology and Psychiatry, 38,* 725–730. doi:10.1111/j.1469-7610.1997.tb01699.x

Charman, T., Ruffman, T., & Clements, W. (2002). Is there a gender difference in false belief development? *Social Development, 11,* 1–10. doi:10.1111/1467-9507.00183

Cutting, A. L., & Dunn, J. (1999). Theory of mind, emotion understanding, language, and family background: Individual differences and interrelations. *Child Development, 70,* 853–865. doi:10.1111/1467-8624.00061

Dunn, J., & Cutting, A. L. (1999). Understanding others, and individual differences in friendship interactions in young children. *Social Development, 8,* 201–209. doi:10.1111/1467-9507.00091

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., . . . Reilly, J. S. (1993). *MacArthur Communicative Development Inventories: User's guide and technical manual*. San Diego, CA: Singular.

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology, 17,* 99–103. doi:10.1037/0012-1649.17.1.99

Flavell, J. H., Flavell, E. R., & Green, F. L. (1983). Development of the appearance-reality distinction. *Cognitive Psychology, 15,* 95–120. doi:10.1016/0010-0285(83)90005-1

Flavell, J. H., Green, F. L., & Flavell, E. R. (1986). Development of knowledge about the appearance–reality distinction. *Monographs of the Society for Research in Child Development, 51*(1, Serial No. 212). doi:10.2307/1165866

Ford, R. M., Driscoll, T., Shum, D., & Macaulay, C. E. (2012). Executive and theory-of-mind contributions to event-based prospective memory in children: Exploring the self-projection hypotheses. *Journal of Experimental Child Psychology, 111,* 468–489. doi:10.1016/j.jecp.2011.10.006

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and appearance-reality distinction. *Child Development, 59,* 26–37. doi:10.2307/1130386

Grant, C. M., Grayson, A., & Boucher, J. (2001). Using tests of false belief with children with autism: How valid and reliable are they? *Autism, 5,* 135–145. doi:10.1177/1362361301005002004

Harris, P. L. (2006). Social cognition. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology: Vol. 2. Cognition, perception, and language* (pp. 811–858). New York, NY: Wiley.

Harris, P. L., Donnelly, K., Guz, G. R., & Pitt-Watson, R. (1986). Children's understanding of the distinction between real and apparent emotion. *Child Development, 57,* 895–909. doi:10.2307/1130366

Hongwanishkul, D., Happaney, K. R., Lee, W. S., & Zelazo, P. D. (2005). Assessment of hot and cool executive function in young children: Age-related changes and individual differences. *Developmental Neuropsychology, 28,* 617–644. doi:10.1207/s15326942dn2802_4

Hughes, C. (1998). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology, 16,* 233–253. doi:10.1111/j.2044-835X.1998.tb00921.x

Hughes, C., Adlam, A., Happé, F., Jackson, J., Taylor, A., & Caspi, A. (2000). Good test–retest reliability for standard and advanced false-belief tasks across a wide range of abilities. *Journal of Child Psychology and Psychiatry, 41,* 483–490. doi:10.1111/1469-7610.00633

Hughes, C., & Ensor, R. (2007). Executive function and theory of mind: Predictive relations from ages 2 to 4. *Developmental Psychology, 43,* 1447–1459. doi:10.1037/0012-1649.43.6.1447

Kochanska, G., Murray, K., Jacques, T. Y., Koenig, A. L., & Vandegeest, K. A. (1996). Inhibitory control in young children and its role in emerging internalization. *Child Development, 67,* 490–507. doi:10.1111/j.1467-8624.1996.tb01747.x

Lalonde, C. E., & Chandler, M. J. (1995). False belief understanding goes to school: On the social–emotional consequences of coming early or late to a first theory of mind. *Cognition & Emotion, 9,* 167–185. doi:10.1080/02699939508409007

Liddle, B., & Nettle, D. (2006). Higher order theory of mind and social competence in school-age children. *Journal of Cultural and Evolutionary Psychology, 4,* 231–244. doi:10.1556/JCEP.4.2006.3-4.3

Mahy, C. E. V., & Moses, L. J. (2011). Executive functioning and prospective memory in young children. *Cognitive Development, 26,* 269–281. doi:10.1016/j.cogdev.2011.06.002

Mahy, C. E. V., & Moses, L. J. (in press). The effect of retention interval task difficulty on children's prospective memory: Testing the intention monitoring hypothesis. *Journal of Cognition and Development*.

Mayes, L. C., Klin, A., Tercyak, K. P., Cicchetti, D. V., & Cohen, D. J. (1996). Test–retest reliability for false-belief tasks. *Journal of Child Psychology and Psychiatry, 37,* 313–319. doi:10.1111/j.1469-7610.1996.tb01408.x

Merrell, K. W. (1999). *Behavioral, social, and emotional assessment of children and adolescents*. Mahwah, NJ: Erlbaum.

Miller, S. A. (1986). Parents' beliefs about their children's cognitive abilities. *Developmental Psychology, 22,* 276–284. doi:10.1037/0012-1649.22.2.276

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. doi:10.1111/j.1467-8624.2007.01018.x

Moore, C., Jarrold, C., Russell, J., Lumb, A., Sapp, F., & MacCalIum, F. (1995). Conflicting desire and the child's theory of mind. *Cognitive Development, 10,* 467–482. doi:10.1016/0885-2014(95)90023-3

Moses, L. J., & Flavell, J. H. (1990). Inferring false beliefs from actions and reactions. *Child Development, 61,* 929–945. doi:10.2307/1130866

Myles, B. S., Bock, S. J., & Simpson, R. L. (2001). *Asperger's Syndrome Diagnostic Scale*. Austin, TX: Pro–Ed.

Newton, P., Reddy, V., & Bull, R. (2000). Children's everyday deception and performance on false-belief tasks. *British Journal of Developmental Psychology, 18,* 297–317. doi:10.1348/026151000165706

O'Neill, D. K. (2007). The Language Use Inventory for young children: A parent-report measure of pragmatic language development for 18- to 47-month-old children. *Journal of Speech, Language, and Hearing Research, 50,* 214–228. doi:10.1044/1092-4388(2007/017)

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology, 5,* 125–137. doi:10.1111/j.2044-835X.1987.tb01048.x

Peskin, J., & Ardino, V. (2003). Representing the mental world in children's social behavior: Playing hide-and-seek and keeping a secret. *Social Development, 12,* 496–512. doi:10.1111/1467-9507.00245

Peterson, C., & Slaughter, V. (2003). Opening windows into the mind: Mothers' preferences for mental state explanations and children's theory of mind. *Cognitive Development, 18,* 399–429. doi:10.1016/S0885-2014(03)00041-8

Pons, F., & Harris, P. (2005). Longitudinal change and longitudinal stability of individual differences in children's emotion understanding. *Cognition & Emotion, 19,* 1158–1174. doi:10.1080/02699930500282108

Raver, C. C., & Leadbeater, B. J. (1993). The problem of the other in research on theory of mind and social development. *Human Development, 36,* 350–362. doi:10.1159/000278223

Ronald, A., Happé, F., Hughes, C., & Plomin, R. (2005). Nice and nasty theory of mind in preschool children: Nature and nurture. *Social Development, 14,* 664–684. doi:10.1111/j.1467-9507.2005.00323.x

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. *Child Development, 72,* 1394–1408. doi:10.1111/1467-8624.00355

Sabbagh, M. A., & Callanan, M. A. (1998). Metarepresentation in action: 3-, 4-, and 5-year-olds' developing theories of mind in parent–child conversations. *Developmental Psychology, 34,* 491–502. doi:10.1037/0012-1649.34.3.491

Schult, C. A. (2002). Children's understanding of the distinction between intentions and desires. *Child Development, 73,* 1727–1747. doi:10.1111/1467-8624.t01-1-00502

Taylor, M., & Carlson, S. M. (1997). The relation between individual differences in fantasy and theory of mind. *Child Development, 68,* 436–455. doi:10.1111/j.1467-8624.1997.tb01950.x

Taylor, M., Cartwright, B. S., & Bowden, T. (1991). Perspective taking and theory of mind: Do children predict interpretive diversity as a function of differences in observers' knowledge? *Child Development, 62,* 1334–1351. doi:10.1111/j.1467-8624.1991.tb01609.x

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children,* 3rd ed. (WISC–III). San Antonio, TX: Psychological Corporation.

Wellman, H. M. (2002). Understanding the psychological world: Developing a theory of mind. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 167–187). Oxford, United Kingdom: Blackwell. doi:10.1002/9780470996652.ch8

Wellman, H. M., Harris, P. L., Banerjee, M., & Sinclair, A. (1995). Early understanding of emotion: Evidence from natural language. *Cognition & Emotion, 9,* 117–149. doi:10.1080/02699939508409005

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75,* 523–541. doi:10.1111/j.1467-8624.2004.00691.x

Wellman, H. M., Phillips, A. T., & Rodriguez, T. (2000). Young children's understanding of perception, desire, and emotion. *Child Development, 71,* 895–912. doi:10.1111/1467-8624.00198

Wetherby, A. M., & Prizant, B. M. (1993). *Communication and Symbolic Behavior Scales: Normed edition*. Baltimore, MD: Brookes.

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.

Appendix

Final Items From the CSUS Broken Down by Mental States

Items marked with an asterisk are found in both the full and short-form scales. Reverse items are marked by (R). The complete version of the Children's Social Understanding Scale (CSUS) exactly as given to parents can be found in Appendix S1 in the online supplement.

Belief

*1. Understands that telling lies can mislead other people.
*2. Talks about how her/his beliefs have changed over time (e.g., "I used to think that drinking from a cup is hard; now I think it's easy").
*3. Talks about people's mistaken beliefs (e.g., "He thought it was a dog, but it was really a cat"; "I thought mommy was coming, but it was really daddy").
4. Tries to persuade others that their point of view is incorrect.
5. Is good at playing tricks on others (e.g., acts as if the cookie jar is empty when really it's full).
6. Talks about what people think or believe (e.g., "I think it's raining"; "He thinks it's bedtime").
7. Talks about differences between his/her beliefs and someone else's (e.g., "You think it's a shark, but I think it is a dolphin").

Knowledge

*1. Realizes that experts are more knowledgeable than others in their specialty (e.g., doctors know more than others about treating illness).
*2. Uses words that express uncertainty (e.g., "We might go to the park"; "Maybe my shoes are outside").
*3. Is good at playing "hide and seek" (e.g., is hard to find, doesn't make give-away noises).
4. Can tell you how she/he found out about things (e.g., "Sally told me about it"; "I saw it happen at the park"; "I heard it on the radio").
5. Is good at explaining things to younger children.
6. Talks about what people know or don't know (e.g., "I know who it is"; "He doesn't know where his ball is").
7. Talks about teaching and learning (e.g., "My dad taught me how to play that game"; "I learned that song at day care").

Perception

*1. Talks about the difference between the way things look and how they really are (e.g., "It looks like a snake, but it's really a lizard").
*2. When talking on the phone behaves as if the listener can actually see her/him (e.g., assumes that the listener knows what she/he is wearing). (R)
*3. Is good at directing people's attention (e.g., points at things to get others to look at them).
4. Thinks that you can still see an object even if you're looking in the opposite direction. (R)
5. Thinks that she/he cannot be seen if her/his eyes are closed. (R)
6. Talks about what people see or hear (e.g., "I see a duck"; "She hears a train coming").
7. Tells lies that are really easy to discover (e.g., says that she/he didn't eat a cookie when there is chocolate all over her/his face). (R)

Desire

*1. Talks about the difference between what people want and what they actually get (e.g., "She wanted a puppy, but she got a kitten").
*2. Takes into account what others want (e.g., takes turns, shares toys, compromises with other children regarding which game to play, etc.).
*3. Talks about differences in what people like or want (e.g., "You like coffee, but I like juice").
4. Understands that wishes don't always come true.
5. Understands that just because you want something doesn't mean you really need it.
6. Talks about what people like or want (e.g., "He likes cookies"; "She wants to go home").
7. Recognizes that if a person wants something, that person will probably try to get it.

Intention

*1. Talks about the difference between intentions and outcomes (e.g., "He tried to open the door, but it was locked").
*2. Has trouble figuring out whether you are being serious or just joking. (R)
*3. Understands that hurting others on purpose is worse than hurting others accidentally.
4. Understands the difference between doing something intentionally and doing it by mistake (e.g., someone deliberately taking a toy vs. taking it by mistake).
5. Understands when she/he is being teased or made fun of.
6. Talks about people's intentions (e.g., "He did it on purpose"; "I didn't mean to spill it"; "She's trying to catch the kitten").
7. Understands that people can perform the same action for different reasons (e.g. throwing a ball could be done with the intention of playing a game vs. with the intention of hurting someone).

Emotion

*1. Understands that different people can have different feelings about the same thing (e.g., one child likes a dog, but another child is scared of it).
*2. When given an undesirable gift, pretends to like it so as not to hurt the other person's feelings.
*3. Talks about conflicting emotions (e.g., "I'm happy to go on vacation, but I'm sad about leaving friends behind").
4. Has difficulty figuring out how you feel from your tone of voice or from your facial expressions of emotions (e.g., has trouble telling the difference between an angry and a sad voice or face). (R)
5. Realizes that if she/he does something bad, others may get mad.
6. Talks about how people feel (e.g., "I am happy"; "She is angry").
7. Tries to understand the emotions of other people (e.g., wants to know why you are crying).

Received May 6, 2011
Revision received July 5, 2014
Accepted July 21, 2014 ■