Direct Assessments of Social Skills can Complement Teacher Ratings in Predicting Children’s Academic Achievement

In press at *Journal of Cognition and Development*

Katharine E. Scott,1,2 Rachel Ann King,1,3 Aaron Cochrane,1,4 Charles Kalish,1 & Kristin Shutts1

1University of Wisconsin-Madison
2Wake Forest University
3University of Chicago
4Brown University

Author’s notes:

All of the tasks and stimuli, R scripts, supplemental analyses, and data on the behavioral tasks for this project can be found on OSF. District-provided individual-level reports are not openly available given the restrictions of our data use agreement. Address correspondence to Katharine E. Scott (Email: kscott@wfu.edu)

Author contributions: C. Kalish and K. Shutts conceptualized this research project. C. Kalish, K. Shutts, and R. King created the study materials. R. King collected the data for the study. K. Scott and A. Cochrane conducted data analyses. K. Scott wrote the manuscript. All authors provided feedback on the manuscript and approved of the final submission.

Funding: This research was supported by the Madison Education Partnership and conducted in collaboration with the Madison Metropolitan School District (Madison, WI).
Abstract

The present research evaluated whether behavioral tasks (“direct assessments”) commonly used to assess young children’s social cognitive development in laboratory studies could have utility for measuring and predicting U.S. children’s outcomes in educational contexts. To do so, children \( (N = 95; \text{49 boys, 46 girls; 41.05\% White, 16.84\% Hispanic, 14.74\% Black, 13.68\% Asian, 11.58\% Multiracial, 2.11\% American Indian/Alaska Native}) \) in a publicly funded pre-kindergarten (4K) program in the United States completed 9 direct assessments that capture important skills in early childhood (e.g., task switching, group conformity preference, theory of mind). The school district also provided children’s 4K and kindergarten grades (assessed through teacher ratings of children) to evaluate whether direct assessments had additional explanatory power over-and-above existing metrics of children’s aptitude. Across the direct assessments, children’s group conformity preferences (i.e., the extent to which children preferred members of the same group to behave in the same way) were most reliably correlated with concurrent (4K) and predictive of future (kindergarten) grades, even when controlling for teacher ratings of children’s concurrent performance. Interestingly, teacher ratings of children on each assessment loaded onto a single factor despite the intention to capture theoretically distinct components of children’s school performance. Discussion focuses on the implications of direct assessments in educational contexts and critical areas for future research at the interface of psychology and education.

*Keywords*: children, social cognition, preschool, academic achievement
Direct Assessments of Social Skills can Complement Teacher Ratings in Predicting Children’s Academic Achievement

The development of social cognitive skills (also called noncognitive skills, character skills, social and emotional skills, and 21st century skills) is one of the most critical tasks of early childhood (Duckworth & Yeager, 2015; Heckman & Rubinstein, 2001). Social cognitive skills in childhood predict a wide variety of important outcomes including future academic achievement, labor market success, health, and criminality (Borghans et al., 2008; Duckworth & Yeager, 2015; Durlak et al., 2011; Gresham, 2016; Heckman & Kautz, 2013). Partially in response to such patterns, educators have increasingly prioritized the importance of fostering students’ social and emotional learning (e.g., CASEL; Graczyk et al., 2000), though social skills have long been a focus of schooling, especially in early childhood (Cartledge & Milburn, 1978).

A critical requirement for fostering social and emotional learning is a clear sense of just what skills and abilities are involved in such learning. Reviewing the literature reveals that there is little consensus concerning best practices for measuring social cognitive skills in early childhood—either at school entry or over time (Duckworth & Yeager, 2015). What skills are important to evaluate and how should they be measured? As reviewed in detail below, many efforts to assess children’s social cognitive skills rely on teacher ratings, with limited efforts to measure children’s social cognitive skills directly. Although teacher ratings can be valuable, they may also have important limitations. The goal of the present research was to investigate whether behavioral tasks commonly used to assess young children’s social cognitive development in laboratory studies could have utility for measuring and predicting U.S. children’s outcomes in educational contexts. Such tasks could potentially supplement teacher ratings, for example, by identifying components of social and emotional learning not well captured by ratings.
Measuring Children’s Social Cognitive Skills

Teacher Ratings

In early childhood education settings, teacher ratings are broadly deployed to evaluate children’s social skills (Gresham, 2016). For example, to assess children’s skills in the domain of emotion, a teacher might be asked to rate on a scale of 1 to 3 how good different children are at reading and responding appropriately to other people’s emotions—and, to track changes over the year, a teacher might be asked to provide such ratings at multiple timepoints (for a review, see Duckworth & Yeager, 2015). Teacher ratings have multiple benefits that contribute to their popularity: 1) they are inexpensive; 2) they can be done relatively quickly; 3) teachers know their students well and see them daily in different contexts; and 4) teachers’ ratings of academic performance are predictive of children’s future academic performance (Hecht & Greenfield, 2001; Südkamp et al., 2012; Teisl et al., 2001). For example, Teisl and colleagues (2001) found that teachers’ ratings of children’s kindergarten performance in math, reading, effort, behavior, and happiness were highly predictive of which children were underachieving in 1st grade. In a meta-analysis, Südkamp and colleagues (2012) found that teachers’ judgments of students’ academic achievement showed a moderate correlation with children’s actual academic performance.

Despite potential advantages of teacher ratings, the predictive ability of teachers’ ratings is certainly not perfect, and there are a number of limitations that bring into question their utility. One core issue with teacher ratings is that carefully evaluating individual students is difficult, especially when schools, at least those in the United States, have an average 10:1 teacher-to-student ratio at the preschool level (NAEYC, 2018). The reliance on a teacher’s ability to recall past behaviors for many different students and integrate memories into a summative evaluation
reduces the ability to sensitively measure change over time (Bowman, 2010; Duckworth & Yeager, 2015). Additionally, when completing assessments of students, teachers must determine a frame of reference for such assessments. However, the frame of reference will systematically differ across each teacher, providing differential meaning to scores from different teachers (Heine et al., 2002).

The social psychological literature provides additional reasons to doubt the objectivity of teacher ratings. First, people believe that they are very consistent over time (consistency bias), but research shows that people behave very differently across situations and time (Mischel, 2009; Podsakoff et al., 2003). Teachers may believe that they are using the same criteria to evaluate children over time but may be responding to items differently without recognizing how their perspectives have changed. Second, teachers’ evaluations of students on different dimensions are likely influenced by a global evaluation of the student (i.e., “halo effect;” Abikoff et al., 1993). If a teacher believes that a child is a “good student,” the teacher may exaggerate that student’s skills in various areas to match their own global perception of the student. Third, in preschool settings, there is a particularly strong focus on the importance of executive function for school readiness (Fitzpatrick et al., 2014; Pellicano et al., 2017), so teachers may not be carefully attuned to other social cognitive skills (e.g., children’s ability to evaluate information from different sources, VanderBorght & Jaswal, 2009; children’s capacity to reason about social relationships, Brey & Shutts, 2015).

Finally, like all members of our society, teachers are susceptible to social biases when evaluating students. For example, when considering identical disruptive behavior from Black compared to White students, teachers rate Black students’ behavior as more troubling and more deserving of disciplinary action than White students’ parallel behavior (Okonofua & Eberhardt,
PRESCHOOL SOCIAL COGNITION

2015; Zimmermann & Kao, 2019). Similarly, teachers have higher expectations of Asian and White students than of Black and Latinx students with comparable academic records (McKown & Weinstein, 2008). Similar biases have also been shown to exist due to other social identities such as social class (Darley & Gross, 1983), gender (Terrier, 2020), and ability (Campbell et al., 2003). Taken together, these findings suggest that teacher ratings of students are not always objective and can be influenced by stereotypes associated with students’ social identities.

The issues we have identified with teacher ratings are potentially why in other domains (e.g., math, reading), teacher ratings are not considered sufficient for assessing children’s skills. In these more traditional academic domains, research has identified components that underlie skilled performance. For example, myriad studies reveal that alphabet awareness and phonological awareness are key components for predicting early literacy (Bradley & Bryant, 1983; Neuman & Dickinson, 2001; Townsend & Konold, 2010; Whitehurst & Lonigan, 2001). These specific components may be challenging for a teacher to identify or isolate from observing a child’s behavior, but carefully crafted assessments can directly measure children’s aptitude. Indeed, there are reliable tests for assessing alphabet awareness (e.g., Test of Early Reading Achievement; Reid et al., 2001) and phonological awareness (e.g., Phonological Awareness Literacy Screen for Preschool; Townsend & Konold, 2010) in early childhood—and these assessments predict concurrent and future literacy (Adams, 1990; Lonigan, 2006; Strickland & Shanahan, 2004; Townsend & Konold, 2010).

In the domain of social cognition, developmental scientists have similarly created direct assessments (also called performance tasks; Duckworth & Yeager, 2015; McKown et al., 2013). These direct assessments capture discrete components of children’s social skills (e.g., emotion perception, Denham, 2006; trust in testimony, VanderBorght & Jaswal, 2009). However,
assessments used in this research literature are not designed to be psychometrically valid measures of individual differences. Thus, their utility in educational setting is an open question. We turn next to describing such tasks.

**Direct Assessments**

Direct assessments typically present children with short, controlled tasks to perform. To illustrate, consider a study Brey & Shutts (2015) designed to measure children’s apprehension of power differences between people. Understanding social power is a critical social skill as it allows individuals to behave appropriately in social interactions and make predictions about other people’s behaviors (Gülgöz & Gelman, 2017; Mast & Hall, 2004; Terrizzi et al., 2019). To study the early emergence of power detection, Brey and Shutts (2015) presented three- to six-year-old children with multiple trials featuring highly controlled displays; each trial featured standardized pictures of people displaying different power cues (such as expansive posture). Children were prompted to identify who was “in charge” across multiple trials. Each child’s responses were coded and summed, resulting in a performance score (for similar measures, see: Castelain et al., 2016; Charafeddine et al., 2015; Enright et al., 2020; Gülgöz & Gelman, 2017; Lourenco et al., 2016; Terrizzi et al., 2019).

Direct assessments, such as the one described above, are typically implemented to measure group-level characteristics. For example, Brey and Shutts’ (2015) measure of social power was used to determine at what age children become attentive to social power cues. The authors concluded that, on average, 5–6-year-old children, but not 3–4-year-old children, were attentive to nonverbal cues of social power. Within each age group, however, there was substantial variability in children’s responses, suggesting that children’s attentiveness to social power may be influenced by other factors in addition to age and that by age 3–4 years, some
children can use nonverbal cues to detect social power (see also Terrizzi et al., 2019). Given variability in children’s responses, it is possible that direct assessments could serve as useful individual difference measures. In other words, the variability within age groups could reflect meaningful individual differences in early emerging social cognitive skills—however, this proposition has yet to be examined.

Using direct assessments to evaluate individual differences in children’s social cognition comes with some clear advantages: In particular, direct assessments can overcome many of the concerns raised earlier about teacher ratings, such as consistency bias, halo effects, and subjectivity in assessment. At the same time, there are potential limitations to using direct assessments to evaluate children’s social skills in educational contexts. For example, the psychometric properties (e.g., test-rest reliability) of many common laboratory measures of social cognition have not been established. And, most notably, administering direct assessments can be costly because they require one-on-one time between a child and an evaluator. As such, direct assessments should only be implemented if they can provide additional individual-level information above and beyond teacher ratings.

**Integrating Approaches for Academic Evaluation**

Although there have been calls for developmental scientists to be more involved in academic evaluation, to date, academic evaluation makes little contact with the literature on children’s social cognition (Hirsh-Pasek et al., 2005). In the present work, we propose that one potential way to overcome limitations associated with both teacher ratings and direct assessments is to integrate approaches and determine the extent to which direct assessments could complement teacher ratings in predicting children’s academic achievement.
Of course, there are many ways to select direct assessments that capture individual differences in children’s social cognition. In the present study, we focus on a set of constructs that are both widely studied in the experimental literature, and that are plausibly relevant to a preschool-age classroom setting. We make no claims that the set selected is comprehensive in covering all aspects of social cognition. Nor are there specific predictions about the relations between different assessments. Notably, although these assessments target constructs widely studied in the research literature, in most cases links to academic performance or teacher ratings of social cognition are unknown. We ask how such measures converge with teacher ratings, capture individual variability in children’s social cognition, and relate to children’s concurrent and future academic aptitude.

**Present Research**

We obtained data from multiple sources to evaluate whether teacher ratings could be supplemented or improved by adding direct assessments of children’s social cognition. See Table 1 for an overview of measures. First, we included direct measures that: (a) had been used previously in educational contexts, and (b) had fairly standardized methods for measuring the constructs. These included two measures of children’s executive functions (task switching: Frye et al., 1995 and Zelazo, 2006; working memory: Simon, 1990) and two measures of children’s apprehension of others’ minds (emotion reading: Parker et al., 2013; theory of mind: Wellman & Liu, 2004). Developmental scientists and educators alike have long emphasized the importance of executive functions for children’s learning and school success; such skills support performance on a wide variety of tasks including those that require adjusting to rule changes and remembering instructions (for a review, see Diamond, 2013). Similarly, developmental scientists and educators have noted the role of understanding others’ minds in supporting children’s
learning and school success. For example, learning information from others (e.g., one’s teacher) requires appreciating another person’s perspective and interpreting their social gestures (see Wellman & Lagattuta, 2014, for discussion).

Second, we included direct measures that met the following criteria: (a) constructs that have been studied by multiple investigators and are familiar contemporary paradigms in research on young children’s social cognitive skills; (b) designed to capture components of social cognition thought to be fundamental to guiding children’s apprehension of, and behavior in, the social world relevant to the classroom; (c) reveal improvements with child age, suggesting that higher scores are indicative of more advanced performance; and (d) capture capabilities that should enhance children’s classroom success.¹

Accordingly, we included two direct measures probing children’s apprehension of social affiliation (Argyle & Dean, 1965; Nurmsoo et al., 2012; Over & Carpenter, 2015) and social power (Brey & Shutts, 2015; Charafeddine et al., 2015; Chudek et al., 2012; Terrizzi et al., 2019). Being successful in social settings (including school) benefits from a capacity to detect how people are connected to one another (for a review, see Platten et al., 2010). For example, understanding cues about who is "in charge" could be helpful for children determining that adults, namely teachers, are in charge in the classroom and that they should do what the teachers request. We also included a measure of children’s reproduction of an adult’s demonstration (Kenward et al., 2011; Keupp et al., 2013) because learning from this kind of observation is critical in both formal and informal contexts (e.g., imitating, following directions; Bjorklund,

¹ We also asked children two simple questions about their liking of school and friends in school. The results from these questions are not included in this manuscript due to lack of variability in children’s responses. Due to an interest from our partnering schools, children also completed measures that assessed social preferences and group-based stereotyping. However, these measures were distinct from the other measures as they do not capture basic social skills and as such, are not included in this manuscript. Measurement details and data for these measures are available in supplemental materials.
2022; Legare & Nielsen, 2015). We further included a task of children’s evaluation of groups conformity (Roberts et al., 2017) because performing well in a classroom setting often involves behaving as others are behaving (e.g., learning group norms; see Veenstra, 2022). Finally, we included a measure of children’s trust in testimony (VanderBorgh & Jaswal, 2009), as a key component of learning new information is discerning which informants are most knowledgeable in particular domains (see Harris, 2012 and Lutz & Keil, 2002 for discussions).

Current practice in social cognition typically involves developing unique measures to address specific research questions. For example, there is no single standard “selective trust” battery, but rather a family of related paradigms. Although we lack consensus about how best to measure social skills in laboratory settings, we selected measures that have been frequently used by other researchers to capture common constructs of interest in the field. In Table 1, we list the main source(s) of inspiration for each measure we included. In many cases, the particular measure included in the present study was an adaptation of a measure (or measures) from the literature. For example, the social power measure in the current study assessed children’s attention to social power as depicted by posture (Brey & Shutts, 2015), body size (Charafeddine et al., 2015; Terrizzi et al., 2018), and audience attention (Chudek et al., 2012). Each of these components was inspired by different original tasks that assessed whether children could identify who was “in charge” based on visual cues, but we developed our own stimuli for each task to create a single cohesive measure. As another example, the theory of mind measure was drawn from Wellman & Liu, 2004, but we adapted that task for presentation on a computer (e.g., showing images of children rather than dolls) to increase feasibility of data collection. We briefly note divergence from the original measures in the method section and provide more details in supplementary materials on OSF.
In addition to direct assessments from children, we obtained the school district’s evaluations of children’s prekindergarten (4K) and kindergarten achievement (report cards) and each child’s scores on a nationally utilized assessment of children’s 4K achievement mandated by the school district and administered as a regular part of instruction (*Teaching Strategies GOLD®*; for prior evaluations of the GOLD, see Qiu et al., 2021; Russo et al., 2019; Vitiello & Williford, 2021). Finally, we asked teachers to rate individual students on the skills measured by each direct assessment—i.e., predict each child’s performance on each assessment—to evaluate whether this could serve as a less resource-intensive proxy for directly measuring children’s social cognitive skills.

**Table 1**

*Overview of Study Measures*

<table>
<thead>
<tr>
<th>Assessment Type</th>
<th>Description</th>
<th>Response Scale (z-scored for analyses)</th>
<th>Source(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Direct Assessments Used Previously in Educational Contexts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Executive function: Task switching</td>
<td>Sort cards according to color or shape, then switch</td>
<td>1 point per correct sorting after switch; possible range 0–8</td>
<td>Frye et al., 1995; Zelazo, 2006</td>
</tr>
<tr>
<td>Executive function: Working memory</td>
<td>Locate color in grid after color disappeared</td>
<td>1 point per correct selection; possible range 0–infinity</td>
<td>Simon, 1990</td>
</tr>
<tr>
<td>Emotion recognition</td>
<td>Identify emotion displayed on face</td>
<td>1 point per correct identification; possible range 0–6</td>
<td>Parker et al., 2013</td>
</tr>
<tr>
<td>Theory of mind</td>
<td>Identify others’ mental state/behavior from vignette</td>
<td>1 point per correct identification; possible range 0–7</td>
<td>Wellman &amp; Liu, 2004</td>
</tr>
<tr>
<td><strong>Direct Assessments Measuring Additional Social Cognitive Skills</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social power</td>
<td>Identify who is “in charge” from a picture</td>
<td>1 point per correct identification; possible range 0–12</td>
<td>Brey &amp; Shutts, 2015; Charafeddine et al., 2015; Chudek et al., 2012; Terrizzi et al., 2019</td>
</tr>
<tr>
<td>Affiliation</td>
<td>Identify affiliation based on eye contact, proximity, or imitative movement</td>
<td>1 point per correct identification; possible range 0–12</td>
<td>Argyle &amp; Dean, 1965; Nurmsoo et al., 2012; Over &amp; Carpenter, 2015</td>
</tr>
<tr>
<td>Imitation</td>
<td>Following the actions of a model</td>
<td>Proportion of actions that matched the model; possible range 0–100</td>
<td>Kenward et al., 2011; Keupp et al., 2013</td>
</tr>
<tr>
<td>Group conformity preferences</td>
<td>Indicate approval or disapproval of group-consistent behavior</td>
<td>Very not okay (1) – very okay (4); possible range -12–12</td>
<td>Roberts et al., 2017</td>
</tr>
</tbody>
</table>
To operationalize our question, we asked whether the direct assessments would predict school achievement (report card data) over and above teacher ratings of children’s aptitude. Such improvement in prediction would provide evidence that the assessments are capturing important variability absent in the teacher ratings. To evaluate the utility of direct assessments for measuring children’s school aptitude, we first evaluated the extent to which children’s performance on our direct assessments was related to teacher ratings of children’s concurrent school performance. Next, we assessed the extent to which teacher ratings in 4K predicted future academic achievement in kindergarten and then whether direct assessments added predictive power above and beyond teacher ratings. Importantly, teacher ratings in 4K assess both typical academic domains as well as children’s social skills. When evaluating the predictive power of direct assessments, we attempted to identify a subset of direct assessments that best predict future academic achievement, and thus, would deserve future attention for refinement and implementation in educational contexts. Finally, we evaluated the convergence between teacher predictions for children’s direct assessment scores and children’s actual performance.

**Method**
Participants

We recruited participants through a state-sponsored pre-kindergarten (4K) program in the United States. The program took place the year before kindergarten with primarily 4-year-old children and was hosted within public schools in the Madison Metropolitan School District. In order to participate, the principal of the school, the classroom teacher, and the parent had to provide permission for children to enroll in the study. The tested sample consisted of 95 children (49 boys, 46 girls) across six public schools. According to school-provided records, participants were White (41.05%), Hispanic (16.84%), Black (14.74%), Asian (13.68%), Multiracial (11.58%), and American Indian/Alaskan Native (2.11%). We were given access to demographic information about children’s socioeconomic status through two different metrics: parental education level and free or reduced lunch status. The highest level of education for the most highly educated parent for each child was as follows: graduate school/professional degree (28.42%), four-year college degree (22.11%), some college/technical school degree (16.84%), high school (21.05%), less than high school (6.32%), and did not report (5.26%). Children were categorized as being eligible for free or reduced lunch (45.26%) or not (54.74%).

Procedure, Materials, and Design

Direct Assessments of Children’s Social Skills

We administered direct assessments of children’s social skills ourselves during their 4K year. Children sat with the experimenter in a quiet corner of their classroom or in the hallway of their school. The experimenter administered each of the tasks individually with each child over the course of two 20-minute experimental sessions. Children completed the tasks in one of eight orders. The orders were determined by grouping the tasks into two sets (Set A: emotion perception, affiliation, working memory, task switching; Set B: theory of mind, social power,
imitation, group conformity preferences, trust in testimony). Children completed either Set A or B first, and completed each set in either forward or backward order. Thus, children could complete Set A forwards or backwards, Set B forwards or backwards, or any combination of these orders. One exception to the order differences was that the imitation and conformity preference tasks always occurred with conformity preference subsequent to imitation. Each direct assessment is briefly depicted in Table 1 and described below.

**Executive Function: Task Switching.** Children’s ability to task switch was assessed by having children sort cards according to color and shape (Frye et al., 1995; Zelazo, 2006). Children completed six trials on which they were told to sort based on color and then eight trials where they were asked to sort based on shape. Children were assigned a point for each correctly sorted card in the shape trials, with possible scores ranging from 0 to 8. This procedure was drawn from the Dimensional Change Card Sorting (DCCS) created by Frye and colleagues (1995; for the full protocol, see Zelazo, 2006). Children were

**Executive Function: Working Memory.** Children completed a Simon span memory task on a tablet to measure their working memory (e.g., Bialystok et al., 2004; Martin-Rhee & Bialystok, 2008; Yang et al., 2011). In this task, participants watched different-colored squares appear on a grid in a particular order. When the colors disappeared, participants were asked to tap each square to fill in the grid in the correct order. After successfully completing one grid, a new grid with two additional squares would appear, and the process would commence again. Scores were calculated as the number of correct box selections children made throughout the game. Scores started at 0 and did not have a maximum. We created this task to mimic prior Simon tasks (e.g., Bialystok et al., 2004; Martin-Rhee & Bialystok, 2008; Yang et al., 2011) but
with reduced language demands; we pilot-tested this task with preschool children prior to beginning this study.

**Emotion Recognition.** Children provided open-ended responses identifying the emotional expressions (happy, sad, mad, scared, surprised, disgusted) displayed on six different children’s faces. If children correctly identified the emotion or provided a synonym for the emotion (e.g., for the “scared” face, “afraid” and “frightened” are also correct), they received a score of 1; if children incorrectly identified the emotion or provided no answer, they received a score of 0. Children’s scores could range from 0 to 6. This task was closely modeled after the Children and Adolescents’ Recognition of Emotion (CARE) “Face Recall” task (Parker et al., 2013) but used different emotion stimuli and a consistent trial order.

**Theory of Mind.** Children completed seven brief tasks to evaluate their theory of mind (Wellman & Liu, 2004). For each task, the experimenter described a situation to the children and children were asked a question about a protagonist’s mental state or behavior and a question checking their memory of the situation. Children received a 1 for every task on which they correctly answered about the protagonist’s behavior or belief and a 0 for every incorrect answer. Scores could range from 0 to 7. We followed the procedure from the Wellman and Liu (2004) with minor adjustments to the stimuli in order to ease presentation (e.g., displaying images of props on a computer rather than using live props).

**Social Power.** Children completed 12 trials where they were asked to identify who was “in charge” (Brey & Shutts, 2015; Charafeddine et al., 2015; Chudek et al., 2012; Terrizzi et al., 2019). On each trial, if children accurately identified the person displaying high-power cues, they received a score of a 1, or if they inaccurately identified the person, they received a score of a 0. Scores could range from 0 to 12. This assessment displayed social power cues that had been used
across different papers to assess children’s attention to social power as depicted by posture (Brey & Shutts, 2015), body size (Charafeddine et al., 2015; Terrizzi et al., 2018), and audience attention (Chudek et al., 2012).

**Affiliation.** Children completed 12 trials on which they were asked to identify people who were likely to be friends (Argyle & Dean, 1965; Baldassare, 1978; Liberman & Shaw, 2019; Nurmsoo et al., 2012; Over & Carpenter, 2015). The trials each portrayed three people, two of whom were displaying affiliative cues through eye contact, proximity, or imitative movement. Children received a score of 1 for every correct answer and 0 for every incorrect answer. Scores could range from 0 to 12. As in the social power measure, this measure depicted different affiliation cues that had been used across papers to assess children’s attention to affiliation as depicted by eye contact (Nurmsoo et al., 2012), proximity (Argyle & Dean, 1965; Baldassare, 1978), and imitation (Over & Carpenter (2015).

**Imitation.** Children completed a tablet game to evaluate their tendency to follow a model (Want & Harris, 2002; Whiten et al., 2006). The objective of the game was to combine and mix ingredients to complete a recipe. The experimenter instructed participants to watch her first to “learn how to do it,” then completed the task before allowing the participant to take a turn. Participants could refer to “recipe cards” illustrating the sequence and ingredients used by the model to reduce memory demands of the task. Scores were calculated as the proportion of ingredient choices that matched the experimenter’s demonstration. The tablet game was created for the purpose of this study and has not been used in prior research.

**Group Conformity Preference.** This measure assessed children’s preference for conformity (Roberts et al., 2017). On eight trials, children were shown members of different novel groups who were engaging in different behaviors (e.g., eating chocolate ice cream vs.
eating vanilla ice cream). After learning about the groups, children were shown a new member of one of the groups. The new person was either engaging in the behavior that was consistent or inconsistent with their group identity. Children were asked whether it was okay or not okay that the new person was engaging in that behavior. If children said that it was okay, they were asked whether it was “very okay” or just “pretty okay;” if children said that it was not okay, they were asked whether it was “very not okay” or just “pretty not okay.” Scores on each trial could range from 1 (very not okay) to 4 (very okay). Children’s preference for conformity was calculated by adding children’s scores on trials that displayed group-consistent behavior (where higher scores indicate preference for conformity) and subtracting scores on trials displaying group-inconsistent behavior (where higher scores indicate preference for non-conformity). Final scores could range from -12 to 12, with higher scores indicating greater preference for group conformity. This task was a simplified version of the original measure (Roberts et al., 2017) to focus on children’s norm evaluations in general rather than on the role of different types of language in norm evaluations.

**Trust-in-Testimony.** Children completed 12 trials on which they were given the option to ask an adult or a child a question about food or toys (VanderBorght & Jaswal, 2009). Half of the trials were questions where an adult would be the most logical trusted source and half of the trials were questions where a child would be the trusted source. On each trial, children received a score of 1 if they identified the correct trusted source and a score of 0 if they did not. This task followed the procedure from the original measure (VanderBorght & Jaswal, 2009) but changed the specific questions so both children and adults would be the trusted source of information for some questions in each domain (i.e., toys and food).

**Teacher Ratings**
4K Teacher Predictions of Children’s Performance on Direct Assessments. We created a single-item question to ask teachers about children’s ability in each domain assessed by the direct assessments. For example, teachers were asked “How well does this child determine who likes who (and who doesn’t)?” to evaluate children’s ability to detect affiliative relationships. Teachers responded to each item on a 1–7 Likert scale.

4K Grades. We obtained information on 4K teachers’ ratings of children’s 4K school performance during the spring semester. The school district’s grading system for 4K students is a standardized assessment involving 31 teacher ratings. The different items capture subscales categorized as: academic (“retells a familiar story in sequence”), individual (“persists with self-initiated activities, seeks support as needed”), social (“engages in social interaction and play with peers”), self (“demonstrates self-control”), and physical (“demonstrates balance and strength”). For each item, teachers provided a score from 1 to 4, with 1 being the worst and 4 being the best.

Teaching Strategies GOLD®. The Teaching Strategies GOLD® is a second assessment of children’s 4K aptitude that utilizes teacher ratings. This measure is the most commonly used assessment measure in state-funded pre-kindergarten programs in the United States (Ackerman & Coley, 2012; Heroman et al., 2010), and was mandatory for programs included in our sample. The Teaching Strategies GOLD® consists of 53 ratings provided by a teacher that measure competency on a number of different dimensions: social-emotional (9 items), physical (5 items), language (8 items), cognitive (10 items), literacy (12 items), and mathematics (7 items). For English language learners, teachers also completed evaluations of English language acquisition (2 items), but given the small number of English language learners in our sample, we do not include these items in our evaluation. Additional questions evaluate science and technology, social studies, and the arts without a rating scale. The items evaluating social-emotional learning
focus on understanding, regulating, and expressing emotions, building relationships with others, and behaving appropriately in different situations. The physical items evaluate motor development. The language items address ability to communicate and understand language. The cognitive items evaluate learning, memory, classification, and use of symbols. The literacy items address phonological awareness, writing skills, and comprehension. The mathematics items evaluate emerging number concepts, spatial relations, measurement, and pattern knowledge.

**Kindergarten Grades.** Finally, we obtained children’s kindergarten grades as reported by their kindergarten teachers at the end of the next fall term (i.e., end of December). Children were given a grade in seven categories: language arts, math, science, social studies, music, physical education, and art. We only examined children’s scores in the core academic domains (language arts, math, science, and social studies) because these were all rated by children’s primary teacher whereas the other scores were assessed by a teacher who only interacted with children occasionally (i.e., during weekly specialized classes), and added additional variability due to differences in teacher rater. Scores in each domain ranged from 1 to 4, with 4 being the highest level of competence. It is noteworthy that unlike the 4K assessment, the kindergarten assessments do not explicitly assess social-emotional skills.

**Results**

**Overview of Analyses**

The analyses reported throughout this manuscript are exploratory and non-preregistered. We had the opportunity to collect data in six schools, allowing us to evaluate how children’s social cognitive skills relate to their school performance. Given the novelty of implementing a set of social cognitive measures and the lack of prior data using such an approach, we did not preregister our analysis plan or target sample size; rather, we included all children whose parents
provided informed consent in the study. Results should be used to inform future research; however, we encourage caution in generalizing specific findings beyond the present sample. District-provided individual-level data are not openly available given the restrictions of our data use agreement, but data for the behavioral tasks, details of analyses, and R scripts can be found on OSF.

Excluded Participants and Measures

Participants with significant developmental delays \((n = 1)\) or a large amount of missing data were excluded (i.e., missing over 10 teacher ratings or at least 1 performance task, \(n = 11\)), and then multivariate outliers on performance tasks were evaluated and excluded from analyses using the robust method detailed by Leys et al. (2018, \(n = 5\); robust covariance estimation used a minimum 90\% of cases; rejection \(\alpha = .01\)). The evaluated sample consisted of 79 children, which provides 80\% power to detect an effect size of \(r = 0.309\).

Creation of Composite Scores

We anticipated that teacher ratings for different items on each assessment (4K grades, GOLD scores, kindergarten grades) would hang together and represent an overall assessment of children’s school achievement (Abikoff et al., 1993). To confirm our hypothesis, we conducted three factor analyses (one for 4K grade items, one for GOLD items, and one for kindergarten grade items) to determine the extent to which different components of teacher ratings related to each other. Within each factor analysis, the items loaded onto a single factor (the second factor’s eigenvalues were below 1 in PCA and EFA scree plots including either the 4K subscores, GOLD subscores, or kindergarten subscores; in each of these cases, simulation-based parallel analyses did not support more than one component). Thus, for 4K grades, we calculated an overall average score across the 31 items to encompass children’s overall 4K competency. Similarly, we
calculated an overall aptitude score for each child on the GOLD. Finally, we calculated an overall kindergarten achievement score by averaging scores in language arts, math, science, and social studies.

**Descriptive Information for Composite Scores**

All measures were z-scored for analyses to ensure comparability of estimated coefficients and improve optimization algorithms by transforming all variables to similar scales. As such, all measures had a mean of 0 and a standard deviation of 1. Importantly, none of the measures had a skew over |0.9|, indicating that none of the measures were highly skewed, and all of the measures had a kurtosis under 3, indicating that the results were unlikely to be driven by a few outliers. See Table 2 for a summary of skewness, kurtosis, and a visual depiction of the distribution of scores on each measure.

**Table 2**

Descriptive Information on Study Measures

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Range of scores prior to transformation</th>
<th>Mean (SD) prior to transformation</th>
<th>Skew</th>
<th>Kurtosis</th>
<th>Distribution of z-scores</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Direct Assessments</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task switching</td>
<td>0–8</td>
<td>4.84 (3.53)</td>
<td>-0.43</td>
<td>-1.67</td>
<td></td>
</tr>
<tr>
<td>Working memory</td>
<td>0–74</td>
<td>20.77 (18.33)</td>
<td>0.70</td>
<td>-0.51</td>
<td></td>
</tr>
<tr>
<td>Emotion recognition</td>
<td>0–6</td>
<td>4.09 (1.20)</td>
<td>-0.87</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Theory of mind</td>
<td>0–7</td>
<td>4.05 (1.65)</td>
<td>-0.49</td>
<td>-0.11</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>-----</td>
<td>-------------</td>
<td>-------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>Social power</td>
<td>3–12</td>
<td>8.41 (1.96)</td>
<td>-0.66</td>
<td>-0.27</td>
<td></td>
</tr>
<tr>
<td>Affiliation</td>
<td>4–12</td>
<td>8.24 (2.20)</td>
<td>-0.15</td>
<td>-1.10</td>
<td></td>
</tr>
<tr>
<td>Imitation</td>
<td>0.21–0.79</td>
<td>0.54 (0.14)</td>
<td>-0.21</td>
<td>-0.61</td>
<td></td>
</tr>
<tr>
<td>Conformity preference</td>
<td>-4.25–5.00</td>
<td>1.03 (2.47)</td>
<td>0.01</td>
<td>-1.24</td>
<td></td>
</tr>
<tr>
<td>Trust-in-testimony</td>
<td>0–12</td>
<td>6.94 (1.94)</td>
<td>-0.24</td>
<td>1.16</td>
<td></td>
</tr>
<tr>
<td><strong>Teacher Ratings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4K grades</td>
<td>1.90–3.90</td>
<td>2.84 (0.47)</td>
<td>0.20</td>
<td>-0.43</td>
<td></td>
</tr>
<tr>
<td>Teaching Strategy GOLD®</td>
<td>22.50–79.00</td>
<td>44.74 (15.19)</td>
<td>0.73</td>
<td>-0.40</td>
<td></td>
</tr>
</tbody>
</table>
Relation Between Direct Assessments and 4K Teacher Ratings

We first examined whether direct assessments captured anything distinct from concurrent teacher ratings that teachers already provided as a part of required student assessments (i.e., 4K grades and GOLD scores). An assessment of the degree of correlation between direct assessments and 4K grades revealed some overlapping variance, but a significant amount of independent variance. Several direct assessments were reliably correlated with 4K grades (all assessed using Spearman ρ, df = 77, and α = .05): task switching (ρ = .27), working memory (ρ = .22), social power (ρ = .44), affiliation (ρ = .25), imitation (ρ = .29), and group conformity preference (ρ = .45). Social power and group conformity preference were the most closely related to 4K grades; however, even these correlations were small enough to reduce concerns regarding isomorphism (i.e., measurement of an identical underlying generative process). Using the same approach, we next evaluated overlapping variance between direct assessments and GOLD scores. No reliable correlations were observed (maximum absolute rank correlation was between GOLD scores and social power, ρ = .21).

Finally, we evaluated whether teachers’ ratings were related to children’s direct assessment scores when teachers were asked directly about the domains captured in the direct assessments (i.e., more proximal questions to the direct assessments). Logistically, if teachers could predict children’s direct assessment scores, the implementation of such predictions would be easier and more cost-efficient than administering direct assessments. To evaluate teachers’ ability to predict children’s social skills, we examined the correlations between teachers’
predictions and children’s scores on each direct assessment. See Table 3. Given the weak correlations between teacher predictions and children’s performance, we would not assume that teacher predictions would be an adequate proxy for children’s direct assessments.

Table 3

Correlations between Direct Assessments and Teacher Predictions of Direct Assessment Skills

<table>
<thead>
<tr>
<th>Direct Assessment</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Switching</td>
<td>0.08</td>
</tr>
<tr>
<td>Working Memory</td>
<td>0.22</td>
</tr>
<tr>
<td>Emotion Recognition</td>
<td>0.16</td>
</tr>
<tr>
<td>Theory of Mind</td>
<td>0.11</td>
</tr>
<tr>
<td>Social Power</td>
<td>0.17</td>
</tr>
<tr>
<td>Affiliation</td>
<td>0.20</td>
</tr>
<tr>
<td>Imitation</td>
<td>0.27</td>
</tr>
<tr>
<td>Conformity Preference</td>
<td>0.20</td>
</tr>
<tr>
<td>Trust-in-Testimony</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Across all of the teacher ratings, there is little convergence between children’s scores on the direct assessments and teachers’ ratings of children’s social cognitive skills or academic achievement. One possibility is that the unexplained variance in the correlations is due to residual noise that is not relevant to children’s kindergarten achievement. Another possibility is that the unexplained variance captures signal in children’s social cognitive skills in 4K that would be helpful in predicting kindergarten achievement. We test these competing possibilities.

4K Teacher Ratings Predicting Kindergarten Achievement

Before evaluating the predictive validity of the direct assessments, we first evaluated whether the 4K teacher ratings that were already implemented by the school district (4K grades, GOLD) predicted kindergarten achievement. This analysis revealed that 4K grades and GOLD scores have 12.0% and 2.7% shared variance with children’s kindergarten grades, respectively (corresponding to correlation coefficients of 0.35 and 0.17). Thus, although 4K grades predicted
kindergarten grades, mean GOLD scores did not reliably predict kindergarten achievement. Moreover, although 4K grades were a significant predictor of kindergarten grades, there was ample unexplained variance in children’s kindergarten grades after considering currently implemented metrics of 4K aptitude. Given the unexplained variance, we next evaluated the extent to which direct assessments could help predict children’s kindergarten achievement.

**Predictive Validity of Direct Assessments**

We first evaluated whether direct assessments measured during 4K would predict children’s kindergarten achievement. For this and later analyses we conducted bootstrapped robust linear models with 2000 resamples for each direct assessment (i.e., to estimate confidence intervals). In addition, we calculated the median change in out-of-sample variance explained by each predictor by fitting robust linear models to 2000 randomly-selected subsets of 80% of the data and subsequently assessing the proportional reduction of error in each subsample’s held-out 20% (i.e., out-of-sample delta R-squared; ΔR²_ooos). Across all the direct assessments, the only measures that significantly predicted kindergarten grades were children’s task switching (b = 0.29, 95% CI: [0.09, 0.52], ΔR²_ooos = 0.13) and children’s group conformity preference (b = 0.34, 95% CI: [0.11, 0.563], ΔR²_ooos = 0.15). See Table 4. Using the same analyses but controlling for age and gender yielded very similar results (i.e., task switching and conformity preferences were the only measures reliably related to kindergarten grades), thus for simplicity we report models without these covariates.

Table 4

**Predictive Validity of Direct Assessments on Kindergarten Grades**

<table>
<thead>
<tr>
<th>Direct Assessment</th>
<th>Bivariate predictiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td>Task Switching</td>
<td>0.29</td>
</tr>
</tbody>
</table>
Following the evaluation of each direct assessment independently, we evaluated the extent to which each direct assessment added predictive power for kindergarten grades above and beyond the other direct assessments. We utilized a cross-validation backwards-selection approach where all direct assessments were first entered into a model as main effects predicting kindergarten grades. Direct assessments were removed starting with the predictor with the smallest proportional reduction in out-of-sample error. Tasks were sequentially removed until all remaining predictors had out-of-sample proportional reductions above 0.01. Consistent with the independent effects reported above, the final model included task switching ($b = .21$, 95% CI: [0.01, 0.42], $\Delta R^2_{oos} = 0.05$) and group conformity preference ($b = 0.25$, 95% CI: [0.02, 0.53], $\Delta R^2_{oos} = 0.07$). In addition, imitation scores predicted a small amount of unique variation in kindergarten grades ($b = .16$, 95% CI: [-0.16, 0.44], $\Delta R^2_{oos} = 0.02$).

**Controlling for District-Provided Teacher Ratings**

Next, we evaluated whether direct assessments added predictive validity over and above already-implemented measures of 4K aptitude by conducting analyses controlling for 4K grades and GOLD scores, in turn. Across all the direct assessments, children’s task switching reliably added predictive power for kindergarten grades when controlling for 4K grades ($b = 0.20$, 95% CI: [0.01, 0.44], $\Delta R^2_{oos} = 0.05$) or GOLD scores ($b = 0.26$, 95% CI: [0.06, 0.49], $\Delta R^2_{oos} = 0.11$). See Table 5. Additionally, children’s group conformity preference reliably added predictive power above GOLD scores ($b = 0.31$, 95% CI: [0.04, 0.63], $\Delta R^2_{oos} = 0.14$) but not 4K grades ($b$

<table>
<thead>
<tr>
<th></th>
<th>$b$</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Proportional Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Working Memory</td>
<td>0.05</td>
<td>[-0.16, 0.39]</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>Emotion Recognition</td>
<td>0.10</td>
<td>[-0.09, 0.39]</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>Theory of Mind</td>
<td>0.16</td>
<td>[-0.02, 0.36]</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Social Power</td>
<td>0.12</td>
<td>[-0.08, 0.36]</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Affiliation</td>
<td>0.18</td>
<td>[-0.02, 0.37]</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Imitation</td>
<td>0.22</td>
<td>[-0.01, 0.48]</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>Conformity Preferences</td>
<td>0.34</td>
<td>[0.11, 0.63]</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Trust-in-Testimony</td>
<td>0.19</td>
<td>[0.00, 0.41]</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>
Together, these analyses suggest that the bulk of the direct assessments do not provide additional predictive power over and above teacher ratings, but task switching and group conformity preferences show promising evidence of capturing unique variance in kindergarten grades.

Table 5

Predictive Validity of Direct Assessments on Kindergarten Grades Controlling for Teacher Ratings

<table>
<thead>
<tr>
<th>Direct Assessment</th>
<th>Controlling for 4K Grades</th>
<th>Controlling for GOLD Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>95% CI</td>
</tr>
<tr>
<td>Task switching</td>
<td>0.20</td>
<td>[0.01, 0.44]</td>
</tr>
<tr>
<td>Working memory</td>
<td>-0.05</td>
<td>[-0.26, 0.23]</td>
</tr>
<tr>
<td>Emotion recognition</td>
<td>0.00</td>
<td>[-0.20, 0.26]</td>
</tr>
<tr>
<td>Theory of mind</td>
<td>0.11</td>
<td>[-0.06, 0.30]</td>
</tr>
<tr>
<td>Social power</td>
<td>-0.03</td>
<td>[-0.22, 0.19]</td>
</tr>
<tr>
<td>Affiliation</td>
<td>0.12</td>
<td>[-0.08, 0.32]</td>
</tr>
<tr>
<td>Imitation</td>
<td>0.13</td>
<td>[-0.15, 0.40]</td>
</tr>
<tr>
<td>Conformity preferences</td>
<td>0.23</td>
<td>[-0.04, 0.54]</td>
</tr>
<tr>
<td>Trust-in-testimony</td>
<td>0.10</td>
<td>[-0.10, 0.31]</td>
</tr>
</tbody>
</table>

*Note. All analyses either control for 4K grades or GOLD scores.*

**Discussion**

In the present research, we evaluated whether direct assessments of children’s social cognition are useful complements to existing teacher ratings in U.S. educational contexts. To do so, children completed 9 direct assessments that are thought to reflect important skills in early childhood (e.g., task switching, group conformity preference, theory of mind) according to the experimental child development literature. Additionally, we obtained teacher ratings of children’s 4K and kindergarten performance. Our results reveal important findings about children’s early emerging social skills, teachers’ ratings of children’s early aptitude, and critical areas for future research at the interface of psychology and education.
One of our primary goals in the present study was to evaluate whether direct assessments could be helpful for assessing children’s performance in preschool classrooms. Across our 9 direct assessments, the tasks that were most reliably correlated with concurrent (4K) teacher ratings on the district-implemented 4K grades were children’s group conformity preferences (i.e., the extent to which children preferred members of the same group to behave in the same way) and social power detection (i.e., children’s ability to identify visual cues about who is in charge). Interestingly, none of the direct assessments were correlated with concurrent teacher ratings on the Teaching Strategy GOLD® (Ackerman & Coley, 2012; Heroman et al., 2010). When considering kindergarten academic achievement, task switching and group conformity preferences in 4K significantly predicted children’s kindergarten grades.

The consistent effects of group conformity preference on concurrent and future academic achievement show that this is an important skill as children are beginning formal schooling. Why might this be? Preference for group conformity may be particularly beneficial for young children in early educational settings because they have a great deal to learn about how to behave in school. If children pay attention to normative group behavior and believe that they should align their own behaviors with the group norm, this may ease their integration into the classroom and give teachers a positive impression of children’s school performance because they are aligning with expectations. Indeed, one of the key goals of early education is to teach children the norms of their social groups, to help children understand acceptable behavior, and to help children become effective group members (Dewey, 1997; LeMaster, 2010).

It is worth noting that the group conformity preference measure does not assess how likely a child is to actually follow a group norm or rule. The assessment closest to this aspect of social cognition, the imitation task, was not uniquely correlated with academic achievement or
Rather group conformity preference may reflect a tendency to perceive group norms. In this task, children are presented only with descriptive information—what some people do. It is up to the child to infer the existence of a norm—what people ought to do. This Descriptive-to-Prescriptive tendency is a robust feature of early social cognition (Rakoczy & Schmidt, 2013; Roberts et al., 2021; Schmidt et al., 2016). Our results suggest that individual differences in this tendency may be related to school performance. Children more likely to interpret regularities in the social environment as having normative force may have an easier time adapting to the expectations of the school environment. One of the challenges of early schooling is coming to appreciate that the school context is different from the home context: The flexibility of family life is not always a good model for the institutional structure of school.

Considering the future implications of group conformity preferences is important given the often-negative connotations associated with a preference for group conformity, also sometimes called prescriptive norms or group stereotyping (Cialdini & Goldstein, 2004). For example, when considering adults’ behaviors, people with strong preferences for group conformity unfairly restrict opportunities that are counterstereotypical for members of particular social groups. For example, many people dislike women who behave assertively because this defies their expectation of how women are supposed to behave (for a review, see Heilman, 2012). Although children’s group conformity preference may improve their ability to succeed in early childhood education settings, will such preferences eventually create rigidity when considering normative behavior in other domains? The long-term implications of group-conformity preferences for academic and social behaviors should be a target of future research.

**Teacher Ratings of Children’s Academic and Social Skills**
In addition to evaluating children’s performance on direct assessments, we presented results that focused on teachers’ ratings of children’s performance. Here, we would like to draw attention to two primary findings. First, although teacher ratings on district-implemented 4K grades, the GOLD, and district-implemented kindergarten grades were supposed to capture multiple components of children’s social and academic aptitude, across all three metrics, teachers’ ratings loaded onto a single factor. These results are consistent with a halo effect (Abikoff et al., 1993) in which teachers provide a global rating of each child that overshadows ratings of individual capacities. However, as in other areas (e.g., production vs. comprehension of language), children can be proficient in one area of social cognition while struggling in other areas. A unidimensional understanding of children’s social skills may prevent teachers from identifying areas in which children need particular attention.

Second, even when we asked teachers to predict each child’s ability in the specific domains measured by our direct assessment, teachers’ predictions had very little correspondence with children’s actual performance on direct assessments. This result could be seen as surprising because teachers have a great deal of contact with their students (Poulou, 2017) and a primary goal of 4K is teaching social skills (Cartledge & Milburn, 1978; Graczyk et al., 2000). Alternatively, as discussed in the introduction, it is not trivial to ask teachers—in classrooms with an average 10:1 teacher-to-student ratio—to recall and summarize each students’ specific behaviors over an extended period of time (Bowman, 2010; Duckworth & Yeager, 2015). Teachers’ inability to predict children’s performance in addition to teachers’ unidimensional evaluations of students suggests a different approach may be needed to understand nuanced differences in children’s early social cognition. One potential method would be providing teachers with direct assessments that they could utilize in the classroom to identify children’s
strengths and weaknesses across a wide array of social skills. Another approach could train teachers to recognize behaviors that correspond with particular social skills. Of course, a final possibility is that teachers failed to predict children’s performance on the direct assessments because the direct assessments were not a good measure of children’s aptitude. Future research is needed to evaluate the best ways to measure and to help teachers assess the diversity of children’s social skills.

Measurement Limitations and Future Directions

In addition to revealing interesting insights about young children’s social cognition and teachers’ understanding of children’s social cognition, the present study identified many areas for future research. One important question will be evaluating why most direct assessments of children’s social cognition did not help predict children’s concurrent or future academic performance. Although we do not have conclusive answers to this question, we can speculate on possibilities that should be evaluated in future research. It is possible that children’s social cognitive skills are not very important for their academic performance. However, we strongly doubt this possibility given a plethora of extant research demonstrating the importance of social cognitive skills for success throughout childhood (Borghans et al., 2008; Duckworth & Yeager, 2015; Durlak et al., 2011; Gresham, 2016; Heckman & Kautz, 2013).

Another possibility is that the direct assessments in the present study lack the validity or reliability to capture children’s social cognitive skills. To date, there has been limited attention to psychometric properties of laboratory measures of children’s social cognition. As such, we did not select measures based on measurement validity or reliability. In future research, it will be critical to identify laboratory measures with excellent psychometric properties. Do measures of children’s social cognition capture individual differences in children’s social cognitive skills?
Should we measure children’s social cognitive skills at many different timepoints to enhance measurement stability? How many trials should we use to maximize reliability while preventing fatigue? Following a careful evaluation of psychometric properties, future research should apply the best direct assessments in educational contexts to replicate the present research.

Finally, it is also possible that the direct assessments chosen in the current study are important predictors of future achievement, but that kindergarten grades are not the right criterion variable. The kindergarten grades collected in the current study are—like the 4K grades—teacher ratings of children’s school achievement. As such, it is plausible that kindergarten teachers are susceptible to all the same biases (e.g., confirmation bias, halo effect) as 4K teachers. These limitations could be overcome by using other criterion variables such as children’s social networks throughout development or observations of children’s social skills by a trained external observer. Additionally, the direct assessments may not capture kindergarten achievement but may capture more distal outcomes throughout childhood. Future research should longitudinally evaluate children’s success over many years and should include other metrics of success in addition to teacher allocated grades. Prior research on the benefits of early childhood education has identified myriad outcomes to capture success including decreased incarceration, decreased teenage pregnancy, increased high school graduation rates, and increased labor incomes (e.g., Doyle et al., 2009; García et al., 2019, 2020). Similar research should be conducted on the impact of early emerging social cognition and how particular aspects of social cognition can help children succeed early and throughout development.

**Conclusions**

The present exploratory study provides suggestive evidence that direct assessments—at least of children’s group conformity preferences—could be helpful in predicting U.S. children’s
academic competency. Additionally, there are several questions that are raised about the use of
direct assessments for assessing children’s social cognition both in and outside of educational
contexts and the structure of teachers’ evaluations of children in preschool settings. We hope that
this paper provides a framework and example of the type of evaluations that could shed light on
children’s emerging social cognitive skills and inspires future research evaluating how direct
assessments of social cognitive skills can be useful in educational contexts.
References


https://doi.org/10.1016/j.evolhumbehav.2011.05.005


https://doi.org/10.1207/s15566935eed1701_4


https://doi.org/10.1146/annurev-psych-113011-143750

https://doi.org/10.1016/j.ehb.2009.01.002


https://doi.org/10.1111/j.1467-8624.2010.01564.x


https://doi.org/10.1016/j.learninstruc.2013.11.003


https://doi.org/10.1207/s1532768Xjepc1101_02


https://doi.org/10.1037/a0026344


https://doi.org/10.1016/j.jecp.2013.07.002

https://doi.org/10.1016/j.tics.2015.08.005


School readiness and the transition to kindergarten in the era of accountability (pp. 121–147). Paul H Brookes Publishing.


