Title

Using Assessment to Improve the Accuracy of Teachers Perceptions of Students Academic Competence.

Permalink

https://escholarship.org/uc/item/5r59z3q7

Journal

The Elementary School Journal, 121(4)

ISSN

0013-5984

Authors

Gatlin-Nash, Brandy
Hwang, Jin
Tani, Novell
et al.

Publication Date

2021-06-01

DOI

10.1086/714083

Peer reviewed

Published in final edited form as:
Elem Sch J. 2021 June ; 121(4): 609–634. doi:10.1086/714083.

Using Assessment to Improve the Accuracy of Teachers' Perceptions of Students' Academic Competence

Brandy Gatlin-Nash[1], **Jin Kyoung Hwang**[1], **Novell E. Tani**[2], **Elham Zargar**[1], **Taffeta Star Wood**[1], **Dandan Yang**[1], **Khamia B. Powell**[1], **Carol McDonald Connor**[1]

[1]University of California – Irvine

[2]Florida A&M University

Abstract

Teachers' perceptions of their students' academic skills can affect students' achievement and may be influenced by unrelated student characteristics such as socioeconomic status (SES). In this ad hoc randomized controlled trial, teachers ($n = 28$) were randomly assigned to receive training on using assessment to guide literacy instruction, Assessment-to-Instruction (A2i), or on Math PALS (control). Teachers rated students' ($n = 446$) academic competence. A2i teachers' ratings did not vary by SES, and their ratings correlated more strongly with students' literacy and mathematics assessment scores compared with those of the control teachers. Control teachers generally underestimated lower SES students' academic competence; underestimation was greater at more affluent schools. Teachers' ratings of students' academic competence predicted reading and mathematics outcomes. Thoughtful use of assessments to guide instruction appeared to improve the precision of teachers' ratings of students' academic competence, improve student outcomes, and reduce potential teacher biases about children from higher-poverty families.

A major source of information regarding students' academic performance is teachers' perceptions of their students' academic competence (Eckert & Arbolino, 2005). Teachers provide daily instruction based on their observations and judgments of each student's knowledge base and competencies. Teachers' perceptions, operationalized as teacher ratings, of their students' academic abilities have been shown to have a direct effect on their expectations for their students (Timmermans et al., 2016), and teachers' expectations are regularly conveyed by their instructional behaviors (e.g., Desert et al., 2009). Students deemed to be less academically competent than their peers are typically provided fewer rich learning experiences, often contributing to what is referred to as a self-fulfilling prophecy of low achievement (see Rosenthal & Jacobson, 1968; Rubie-Davies, 2010). Thus, teachers' expectations of student performance, and thereby their perceptions of students' academic competence, may affect the instruction they provide and, in turn, predict students' academic achievement. In this post hoc study, we examined whether providing professional development (PD) to randomly assigned teachers on how to use assessment to guide literacy instruction might improve the accuracy of teachers' ratings of students' academic competence compared with teachers in the control condition. In addition, we

Correspondence concerning this article should be addressed to Brandy Gatlin-Nash, University of California, Irvine. gatlinb@uci.edu.

examined whether various student, teacher, and school-level characteristics might be related to teachers' ratings of students' academic competence within the context of their actual performance.

Teacher Perceptions of Students' Capabilities

Research evidence suggests that teachers' perceptions of academic abilities may be associated with student demographic characteristics. For instance, after measuring achievement among a large sample of diverse kindergarten children, Ready and Wright (2011) concluded that teachers in their study underestimated Black, Hispanic, and language-minority Asian students' language and literacy skills. They reported discrepancies between teachers' ratings and students' performance on achievement measures. In a study with first-grade students conducted by Hinnant et al. (2009), teachers overestimated reading abilities for girls and underestimated reading abilities for boys, noting that teachers' perceptions of students' abilities predicted achievement years later. Speybroeck et al. (2012) found that teachers tended to rate students from more advantaged backgrounds as academically stronger than children from low socioeconomic status (SES) backgrounds, controlling for student race, gender, and prior achievement. Also, teachers in higher SES classrooms have been shown to overestimate students' abilities in literacy, whereas teachers in lower SES classrooms tended to underestimate students' abilities (Ready & Wright, 2011). These findings emphasize the importance of considering contextual SES in addition to individual student SES in studies involving teacher perception of academic competence

Teachers' perceptions of students' social skills and externalized behaviors on academic outcomes and gains also appear to be related to teachers' ratings of students' academic outcomes and gains. In a sample of preschool teachers and their students from low-income backgrounds, Baker et al. (2015) found teachers rated students with stronger social skills and fewer inattentive symptoms as having better preacademic skills than their peers. With social and behavioral ratings and achievement held constant, student race and gender were not associated with teacher perceptions of academic ability. Robinson-Cimpian et al. (2014) conducted a study in which kindergarten through third-grade teachers' perceptions mediated gender gaps in students' mathematics achievement across grade levels. In particular, teachers often conflated their ratings of students' behavior with their ratings of students' mathematics proficiency. Teachers generally underrated girls when considering perceived student behaviors, and viewed boys as better in mathematics than girls with equal achievement and similar behaviors. Thus, teachers' judgments of students' social skills and behaviors can potentially influence their judgments of academic achievement.

Few experimental studies have investigated teacher expectations and their effects on students' academic gains. In their review of studies designed to examine teacher expectations and students' academic outcomes, Rubie-Davies et al. (2015) pointed out the need for randomized controlled trials (RCTs) designed to change teacher beliefs and instructional practices. As the researchers noted, most of the research at the time had purported to address methodological concerns following the seminal Rosenthal and Jacobson (1968) Pygmalion experiment. In the original study, researchers informed teachers that several of their students were "bloomers" destined to thrive academically. At the end of the school year, “bloomers” were reported as having made greater intellectual gains. Most of those studies, however, were descriptive studies, as opposed to RCTs, and several induced false expectations among study participants (i.e., teachers). The intervention-based experiment conducted by Rubie-Davies et al. (2015) is the only study, to the best of our knowledge, that has sought to positively influence student achievement by changing teachers’ expectations about their students’ capabilities. Specifically, teachers randomly assigned to the intervention condition attended workshops aimed at increasing instructional behaviors modeled after teachers who held high expectations for all students. Teachers were taught to emulate behaviors that were utilized by these high-expectation teachers. Control teachers received regular PD provided by their school. The intervention was effective in improving students’ mathematics performance compared with control students but did not lead to significant gains in students’ reading skills. The authors concluded that adjusting non-content-specific instructional practices may affect reading and mathematics development differently. What is not clear is whether teachers’ perceptions of their students differed as a result of the intervention. Although post hoc in nature, the current study extends our current knowledge by measuring differences in treatment and control teachers’ perception of students’ abilities as part of an RCT intervention study, which included non-content-specific aspects (individualized instruction) within the context of two different areas – reading and math.

Assessment-Informed Instruction

The debate on the utility and value of assessment continues among educators, policy makers, and parents. Although educators understand the importance of both formative and summative assessments (e.g., low-stakes, curriculum-embedded measures and high-stakes standardized tests, respectively), the incorporation of either into instructional practices can be difficult for many teachers (Black, 2015; Dunn & Mulvenon, 2009; Roehrig et al., 2008). Black (2015) analyzed several studies where teachers implemented novel formative assessment measures to inform their instructional practices. Research interventions using formative assessment have generally predicted student gains in settings where educators utilize knowledge-based learning approaches that measure students’ acquisition and application of content-based knowledge through assessments (in reading, e.g., Al Otaiba et al., 2014; Connor, Morrison, Fishman, & Schatschneider, 2013; and in mathematics, Burns et al., 2010). Although overall effective in improving student learning, these teaching practices frequently required periods of teacher and student adjustment (Roehrig et al., 2008). When considering summative, often high-stakes assessments, teachers have reported teaching to the test (Popham, 2001) and frequently focusing their attention on students who are most likely to progress (Hunter, 2019). Researchers contend that teachers’ access to summative assessments may have positive effects by improving instructional quality and practices provided as early as prekindergarten and kindergarten (e.g., high-quality, age-appropriate instruction grounded in evidence-based research and benchmark-inclusive curricula; Graue et al., 2017). Others agree, arguing that high-stakes summative and standardized assessments may serve as well for formative purposes (Dunn & Mulvenon, 2009; Wininger, 2005). At the same time, results from such high-stakes assessments may arrive too late to help inform instruction. Whereas assessments can be a means of guiding instruction that is tailored to the learning needs of individual students (Connor, Morrison, Fishman, & Schatschneider, 2013; Moon, 2005), assessments might also serve to increase the accuracy of teachers' perceptions of their students' academic capabilities. Assessments may offer an avenue for improving teachers' understandings of their students' strengths and weaknesses. Thus, the present study was conducted to examine whether an intervention designed to support teachers' use of assessment to guide individualized instruction might also improve the accuracy of their ratings of students' academic competence.

Assessment-to-Instruction Technology

The current study utilizes data from a cluster RCT investigating the impact of assessment-data-driven individualized instruction on students' literacy outcomes using Assessment-to-Instruction (A2i) technology (Connor, 2013). A2i is a web-based teacher professional support system designed to improve teachers' use of assessment data to individualize literacy instruction. Screenshots are provided in Figures A1 – A3. The dynamic forecasting intervention algorithms that are embedded in A2i calculate the amounts (in minutes per day) and types of literacy instruction each individual student should receive to achieve optimal growth in reading. The algorithms use nationally normed standardized test scores in language and literacy to compute recommended amounts of literacy instruction (Connor et al., 2007). Using assessment information, A2i also recommends small flexible learning groups consisting of students with similar learning needs. Research-based lesson plans and activities are available so that it is easy for teachers to incorporate evidence-based materials into their instruction. Online PD resources are also an important component of A2i; information about interpreting assessment scores, videos of master teachers using assessment to guide instruction, and detailed information on individualizing instruction are available online. A2i also allows educators to track students' progress over time through user-friendly graphs. Instead of having teachers interpret assessment scores and provide instructional goals based on personal judgments, A2i facilitates this process by recommending amounts of instruction in the Classroom View (Fig. A2) and specific learning activities through the Lesson Plan features.

Few studies have examined teacher perceptions of emergent readers and students' subsequent academic gains (Baker et al., 2015). The vast majority of literature surrounding teacher perceptions has focused on teacher judgment accuracy (Machts et al., 2016; Meissel et al., 2017; Südkamp et al., 2012, 2017). Studies of teacher perceptions and students' academic gains have relied on noninstructional intervention models (Rubie-Davis et al., 2015). Finally, given the limited number of teacher perception studies conducted in the United States on emergent learners (Baker et al., 2015; Ready & Wright, 2011), the present study adds to the literature by exploring associations between teacher perceptions of academic competence and reading comprehension and mathematics development in first graders. In this post hoc analysis of a cluster RCT, we use the power of random assignment (i.e., teachers randomly assigned within schools) to compare teachers who used A2i technology with control teachers who received PD on an evidence-based mathematics intervention, Math Peer-Assisted Learning Strategies (Math PALS; Fuchs et al., 1997), but obtained only paper reports of test scores. We aim to investigate whether receiving specific training in using assessment to individualize students' instruction, through the

A2i technology, improves teachers' precision in rating students' academic competence and reduces the use of less relevant characteristics such as race, gender, behavior, and SES to inform their decisions. Thus, the following research questions inform this study:

1. Does teachers' participation in the A2i intervention (i.e., access to A2i intervention and PD about personalizing literacy instruction) predict their ratings of students' competence?

    a. To what extent do student-level (i.e., gender, race, behavior, academic skills, National School Lunch Program [NSLP] status) and school-level characteristics predict teachers' ratings of students' academic competence, controlling for teacher participation in the A2i intervention?

2. To what extent do teachers' reported ratings of students' academic competence predict gains in reading comprehension and mathematics from fall to spring, controlling for student-, classroom-, and school-level characteristics?

We hypothesized that with the affordances of A2i technology (assessments to guide individualized instruction, recommending flexible learning groups, providing PD and well-designed graphs of assessment information to the teachers) coupled with PD on personalizing literacy instruction, A2i teachers' ratings of students' academic competence would be more accurate, and the potential bias in judgments based on student characteristics would be reduced compared with control group teachers. Although A2i is a literacy intervention, the essence of A2i is to help teachers understand and interpret students' test scores and apply them to personalize instruction based on their students' needs. We believed that this skill could be transferrable to other content areas. Further, teachers in the control condition also received PD about incorporating students' assessment scores in their instruction (in math; more details about PD are reported in the Method section). Based on the extant research, we hypothesized that teachers' ratings of academic competence would predict students' reading and mathematics skills, controlling for other student characteristics, such as behavior, race, and gender.

Method

Participants

Four hundred forty-six first-grade students from 28 classrooms and their teachers participated in this RCT during the 2008–2009 school year (see Connor, Morrison, Fishman, Crowe, et al., 2013). The students attended five schools from one school district in northern Florida. This school district served ethnically and economically diverse student populations in rural, suburban, and urban communities. In general, 47% of the students were eligible for the free or reduced-price lunch program, or the NSLP, a commonly used indicator of low SES.

At the beginning of the school year, 28 teachers and 446 students (with parent consent) joined the study: 15 teachers and their 255 students were randomly assigned into the A2i treatment group and 13 teachers and their 214 students into the Math PALS control group (described in the Procedures section). Teachers completed the *Social Skills Rating System*

(SSRS; Gresham & Elliott, 1990) for the 446 students included in the study presented here (i.e., SSRS ratings for 23 students were not completed; thus, they were not included in the data analyses; A2i students $n = 245$; control $n = 201$). Missing data analyses suggested that students with missing data tended to have lower fall passage comprehension scores ($M = 434$ compared with 442). There were no other significant differences in any of the other key variables. Representative of the community in which the study took place, 84% of the students were White, 6% were Multiracial, 5% were African American, 3% were Hispanic, 2% were Asian, and 0.7% were Native American. Approximately 46% of the students were boys. Twenty-seven percent of the students qualified for NSLP. Chi-square tests indicated that students were evenly represented as far as demographic characteristics (gender, ethnicity, and NSLP status) across the A2i treatment group and the mathematics control group (X^2s = .41–8.32, $p > .05$).

The 28 participating teachers were all female, with an average of 17 years ($SD = 11$) of teaching experience. All but one identified as White, and one identified as African American. Twenty-eight percent of the participating teachers reported having an MA or MS degree, and 7% had an MEd. Furthermore, 31% held certification in early childhood development, 86% in elementary education, 17% in exceptional student education, and 3% in reading. The five participating schools were located in one district and varied in context from more urban to rural. School percentage of students in the NSLP ranged from 31% to 45%, with a mean of 37%. These data were downloaded from the National Center for Education Statistics (NCES) website the year of the study (https://nces.ed.gov/ccd/schoolsearch/). All schools used the same reading and mathematics core curriculum, Harcourt Trophies and Saxon Math, respectively.

Measures and Procedures

Students' literacy and mathematics skills were assessed individually in a quiet place in the students' school in the fall, winter, and spring quarters during the school year. This study used the fall and spring data.

Literacy assessments.

Students were assessed on their literacy skills using the Woodcock-Johnson III Tests of Achievement (WJ) letter-word identification and passage comprehension subtests (Woodcock et al., 2001). Students' word reading skills were assessed with the letter-word identification subtest. The passage comprehension subtest was used to assess students' reading comprehension skills. This subtest measures students' understanding of what they read using a cloze task. W scores were used in the analysis. W scores are scaled on a single metric and allow for the measurement of students' growth over time (Jaffe, 2009). The scale of W scores is centered on a value of 500, which is equal to the average performance of a typical child age 10 or at the beginning of grade 5. The typical range of W scores for this age range within a test is about 430 – 550 (Woodcock et al., 2001). Reported reliability ranges from .90 to .96 for students 2 – 7 years old for the two assessments.

Mathematics assessment.

Students' mathematics competency was assessed using the applied problems subtest of the Woodcock-Johnson III Tests of Achievement (Woodcock et al., 2001). Applied problems assesses ability to analyze and solve math problems. Again, W scores were used in the analyses. Reported reliability of this subtest is .96 for students 2–7 years old.

Teacher ratings of academic competence, social skills, and problem behaviors.

Teachers' ratings of their students' academic competence, social behaviors, and behavior problems were measured using the SSRS (Gresham & Elliott, 1990), a norm-referenced, multirater assessment tool for children and youth in pre-K through twelfth grade. The SSRS contains 57 items in three measurement areas of academic competence, problem behaviors, and social skills. The teachers were asked to fill out and return this questionnaire for each student in January or February of the school year. This was an appropriate time of the academic year for teachers to rate their students' academic competence because they would have known and taught their students for more than 4 months. Students' academic competence level was measured by comparing the student with other students on a scale from 1 to 5, where 1 represents the lowest 10%, 2 is the next lowest 20%, 3 is the middle 40%, 4 is the next highest 20%, and 5 is the highest 10%. Of note, only four questions in the teacher SSRS academic competence subscale asked teachers to evaluate students' performance in reading and math (e.g., "In reading, how well does this child compare with other students?"). In addition, teachers answered five questions regarding students' overall academic performance, motivation, parental encouragement, intellectual functioning, and classroom behavior (e.g., "The child's overall motivation to succeed academically is"; Gresham & Elliott, 1990). These scores provide a composite score used to gauge teachers' global judgment of students' academic competence.

In the social skills section, both frequency and the importance of each social or interpersonal behavior were rated on a scale (i.e., never, sometimes, or very often for frequency, and not important, important, and critically important for the importance scale). In the problem behavior section, only frequency was included. The raw scores were converted to standard scores (SSs) with a mean of 100 and a standard deviation of 15. Of note, higher academic competence and social skills scores are associated with more positive ratings. For behavior problems, higher scores indicate teachers perceive students to have more serious behavior problems (i.e., low scores indicate good behaviors).

The reliability of SSRS was estimated by internal consistency and test-retest. The average coefficient alpha reliabilities for internal consistency was .90 for social skills, .84 for the problem behaviors, and .95 for the academic competence subscale (Community-University Partnership for the Study of Children, Youth, and Families, 2011). According to the manual (Gresham & Elliott, 1990), test-retest analyses yielded coefficient alphas between .84 and .93 across all three areas for teacher report. Given the post hoc nature of the present study, we used the measures available to us. We recognize that other metrics may be better suited to address some of the study aims; namely, given the non-domain-specific nature of the academic competence subscale, the study findings are limited in scope based on the measures at the researchers' disposal (see Südkamp et al., 2012, and the Discussion section for limitations).

Demographic data.

Student NSLP status, gender, race/ethnicity, and date of birth were obtained through the school, the Florida database (Progress Monitoring Reporting Network), and through questionnaires sent home to parents/guardians in the early spring of the school year. Teachers completed questionnaires, which provided us with information on their years of experience, training, gender, and race/ethnicity. We examined whether students of particular race/ethnic groups were more likely to attend particular schools. We found that students who identified as Black were more likely to attend the school with a lower percentage of students qualifying for the NSLP (i.e., more affluent). Students who identified as White were more likely to attend the school with the highest percentage of students qualifying for NSLP.

Procedures

These data were collected as part of a cluster RCT (Connor, Morrison, Fishman, Crowe, et al., 2013). Teachers and their students were randomly assigned within schools to either the treatment or control condition. Both the treatment and control condition classroom teachers received the same amount of PD, but the focus varied by content area—reading or mathematics—and by engagement in the use of assessment to inform instruction. The rationale for an alternative treatment control rather than a business-as-usual control was based on the request of the school district to have a mathematics treated control group and because one purpose of the study was to examine whether amount of time and type of PD was responsible for previous study reported effects. Thus, the amount and type of PD across both groups was equivalent and only the content of the PD (reading or mathematics) and focus on assessment-driven instruction varied by condition.

The teachers in the control condition delivered business-as-usual instruction during their literacy block but implemented a research-based mathematics intervention (i.e., Math PALS; Fuchs et al., 1997) for their mathematics class periods. For our treated control condition, teachers did not tailor instruction based on student math assessments. The A2i treatment group teachers, on the other hand, were more deeply engaged in the use of assessment to inform the transformation of teaching practice and knowledge about students as individuals. Greater personalization of instruction, as guided by assessment, sets the A2i treatment group apart from the Math PALS treated control condition. The teachers in the A2i condition had access to A2i, where they could easily access students' literacy assessment data, but they did not receive any additional support in mathematics instruction. All teachers received the results of both reading and mathematics assessments. Teachers in the A2i condition accessed reading scores through the A2i technology in graphic and table formats, whereas teachers in the Math PALS condition received paper reports of reading assessments. Teachers in both groups received paper reports of mathematics scores.

PD protocols by condition.

In both the A2i treatment condition and Math PALS treated control condition, teachers were assigned to initial 2-day intensive training sessions at the beginning of their school year. On day one, routines and procedures of the interventions were covered. On day two, more conceptual and content-specific aspects were covered. The initial training sessions for A2i were led by research assistants who were highly trained and fully credentialed. The teachers in the treatment condition received PD on individualizing literacy instruction and how to use A2i in their instruction—which included using assessment to inform instructional decision-making, methods to provide individualized instruction, and classroom management including how to manage and engage in small-group instruction while the rest of the students engaged in learning on their own or with peers. The Math PALS initial 2-day-long training sessions were led by a researcher with extensive experience with the Math PALS intervention. Directed by the Math PALS training manual (Fuchs et al., n.d.), teachers received intensive instructions on how to evaluate students' math abilities based on assessment scores, pair students based on their rank order, and guide students to conduct Math PALS learning activities following the training manual. Working in heterogeneous ability pairs with a higher-ability child and a lower-ability child taking turns in supporting one another, all students were engaged in learning activities to support numeracy as appropriate in first grade, as well as fluency with mathematical equations and more applied mathematics skills. In Math PALS, although the intervention includes curriculum-based measurement assessments for creating the learning dyads, frequent and strategic use of assessment to drive instructional decision-making is not part of the focus. Rather, the scope and sequence of Math PALS is predetermined by the manual.

Training sessions continued in monthly communities of practice meetings (Bos et al., 1999) delivered by research partners from the university conducting the study. Research partners were former teachers and practitioners familiar with classroom teaching as well as the interventions. In the communities of practice meetings, teachers were supported in implementing individualized student instruction in reading (A2i) or peer learning strategies via Math PALS to fidelity. The topic of using assessment in the A2i intervention to personalize literacy instruction via teacher-managed small groups and independent child-managed small groups was discussed in the communities of practice meetings for the A2i condition. For the Math PALS condition, the use of measurement to create dyads as well as alter dyads monthly was discussed. In addition, both conditions received bimonthly, in-class coaching sessions in which research partners from the university who were delivering the monthly communities of practice meetings would conduct in-class PD, which included demonstrating small-group instruction in A2i or monitoring of Math PALS dyads in the treated control condition.

Fidelity.

To measure fidelity of implementation across conditions of treatment (A2i) and treated control (Math PALS), which represent two separate interventions, two separate approaches were taken. Fidelity of implementation for the A2i treatment condition was assessed in two ways: monitoring the use of A2i using the teacher user log information and classroom videotaped observation. More specifically, teachers' log data were automatically recorded and generated by the A2i software; this includes the date and time teachers visited the A2i website, the exact pages they visited, and the time they spent on each page. We evaluated the time teachers spent on managing and viewing the A2i online assessments, planning their lessons in A2i, viewing the teaching resources, and using the discussion board. Videotaped classroom observations were conducted to compare the instructions students received in the two conditions. Trained research assistants coded the amounts and types (e.g., code-focused or meaning-focused; whole class or small group) of literacy instruction.

Given the different nature of the Math PALS intervention, we coded classroom videotape data of both the A2i literacy group and the Math PALS intervention group during their math lessons to compare amounts of time spent in peer-managed activities in math across all classrooms. This was undertaken to ensure fidelity of Math PALS implementation as far as it is a set of activities designed to complement a curriculum with its additional, peer-managed activities. For the sake of examining fidelity of Math PALS in our study, we needed to verify if students in the Math PALS condition engaged in more peer-assisted learning opportunities during math compared with students in the A2i classrooms during their business-as-usual math block with only Saxon Math. We used ANOVA to explore this. Results showed the overall amount of time Math PALS students participated in peer-assisted activities was about 10 minutes daily ($SD = 6.91$). Students in control classrooms spent 6 minutes in math peer-assisted learning opportunities ($SD = 6.52$). Students in the Math PALS condition spent significantly more time working in small, peer-managed groups than did students in control classrooms ($F = 109.22$, $df = 1, 289$, $p < .001$).

Results

All collected data were cleaned and analyzed in SPSS and hierarchical linear modeling (HLM), due to the nested nature of the data, as described in more detail in this section. Descriptive statistics are provided in Table 1 and correlations in Table 2. On average, children’s reading and mathematics skills fell within grade-level expectations but with wide variability. We used *t* tests to examine whether there were any between-group differences in students’ fall WJ assessment scores after the random allocation. We did not find any significant group differences in baseline scores (*t*s = –1.68 to –.69, *p*s < .05). Examination of correlations among teachers’ ratings of academic competence and standardized assessments of literacy and mathematics revealed moderate to strong correlations. Correlations were greater for students in A2i classrooms ($r = .48$–$.73$) than for students in control classrooms ($r = .33$–$.45$).

Because this was a cluster RCT where teachers within schools were randomly assigned to either treatment or control condition, students were nested in classrooms, which were nested in schools. To consider the nested structure of these data, we used HLM (Raudenbush & Bryk, 2002) for the analyses. We first conducted separate three-level models with teachers’ rating of students’ academic competence as the outcome for research question 1 and student passage comprehension and mathematics achievement as the outcomes for research question 2, with students nested in classrooms nested in schools. There was no significant variance at the school level once school-wide NSLP percentage (school % NSLP) was added to the models. To create more parsimonious models, we used two-level models with students nested in classrooms and added the school % NSLP variable at the classroom level. We also trimmed nonsignificant variables. Models (provided in Tables 3–5 and the Appendix) were built systematically starting with an unconditional model, which we used to compute the intraclass correlation (ICC). ICC represents the proportion of variance explained at the classroom level (i.e., between classrooms). Continuous variables were sample grand mean centered except for the SSRS variables, which were classroom mean centered. By grand mean centering, the intercept represents the fitted mean of students with sample average scores. However, because teachers rated students using the SSRS, they likely considered students in the context of their classmates. Therefore, we group (i.e., classroom) mean centered the students' SSRS scores.

Student and School Characteristics Predicting Teacher Rating of Academic Competence

To answer our first research question, we built HLM models with teacher rating of academic competence as the outcome. We created an interaction term to capture the potential three-way interaction between assignment to A2i treatment group, school % NSLP, and student NSLP. For race/ethnicity, we used White as the fixed reference group. We also created achievement gain variables to capture students' mathematics and reading gains over first grade by subtracting the fall W score from the spring W score.

The ICC for the unconditional model was .52. Thus, more than 50% of variability in ratings of students' academic competences was explained by which classroom the student attended. Teachers in the A2i condition tended to align ratings of students' academic competence more closely with assessment results than did control teachers (see Table 2). That is, their ratings correlated more highly with standardized literacy and mathematics assessments. Teachers in the A2i condition also rated their students as more academically competent overall than did control teachers, controlling for student and school characteristics (see Table 3 and Fig. 1). However, this is complicated by the three-way interaction effect—A2i X school % NSLP X student NSLP—on teachers' rating of students' academic competence. Figure 1 shows teachers' ratings of students' academic competence, controlling for actual literacy and mathematics skills (sample grand mean centered), literacy and mathematics gains, and students' rated behavior problems and social skills (classroom mean centered) as a function of student NSLP status and school % NSLP. Overall, teachers in the A2i group judged the academic competence of students who qualified for NSLP as similar to those who did not qualify, regardless of school % NSLP. Moreover, ratings were close to a SS of 100; this is appropriate as the reading and mathematics skills were centered at the grand mean of the sample, which was close to national norms. Teachers in the control classrooms rated students who qualified for NSLP as significantly less academically competent than students who did not qualify for the NSLP, and lower than peers in A2i classrooms. This varied by school context, however, with the greatest gap at the more affluent schools and a reverse gap at higher % NSLP schools. That is, control teachers' ratings of students' academic competence were lower overall and depended on individual students' NSLP status, as well as the school % NSLP, whereas A2i teachers' ratings were higher overall and did not vary significantly by student NSLP status and school % NSLP.

Students' fall literacy and mathematics achievement level and achievement gains over the school year also predicted teachers' rating of academic competence (see Table 3). In addition, as teachers' perceptions of students' behavior problems increased, teachers' rating of academic competence decreased. Plus, teachers generally rated students with stronger social skills as more academically competent than students with weaker social skills. Gender and race/ethnicity did not predict teachers' ratings (see Table A1 for model with race/ethnicity included). Race/ethnicity variables were trimmed from the model to preserve parsimony.

Predictors of Reading Comprehension and Mathematics

Our second research question specifically asked to what extent teachers' ratings of academic competence predicted reading comprehension and mathematics gains (residualized change), controlling for other student, classroom, and school characteristics. We first built a model to examine predictors of students' reading comprehension gains with spring passage comprehension as the outcome variable and then a model with spring mathematics as the outcome. Teacher ratings of academic competence, social skills, and behavior problems SSs, students' gender (girl = 1, boy = 0), student NSLP status (1 = qualified, 0 = not qualified), fall reading and math W scores, and student race/ethnicity were entered at the student level. We created dummy variables for each of the race/ethnic groups (1 = reference group, 0 = all others), with students who were White as the fixed reference group. School % NSLP and treatment condition (A2i = 1, control = 0) were entered at the classroom level. Results are presented in Tables 4 and 5. The ICCs are reported with the tables. The ICC for the unconditional model indicated that approximately 5% of the variability in students' passage comprehension scores was explained by the classroom they attended.

Reading comprehension.

HLM results revealed that teachers' ratings of students' academic competence significantly predicted students' reading comprehension gains (see Table 4). For every 1-point increase in teachers' rating of academic competence, students' passage comprehension scores increased .24 points – a 4.4 W point difference for a one standard deviation increase in teacher ratings (*SD* = 18.18). This translates into an effect size (*d*) for a one standard deviation change in teachers' rating of .44, which is an educationally meaningful effect (Hill et al., 2008). Overall, students in the A2i classrooms achieved gains in reading comprehension scores that were greater than students in the Math PALS control group (*d* = .47). Fall achievement scores in both literacy skills (i.e., letter-word identification and passage comprehension) and mathematics significantly predicted students' spring reading comprehension skills. Notably, students' NSLP status, school-wide % NSLP, teachers' ratings of social skills, and reported behavior problems did not predict gains in reading comprehension (although they did predict teacher ratings of academic competence). Students' race/ethnicity and gender did not significantly predict reading comprehension gains.

Mathematics.

When we considered students' mathematics outcomes (see Table 5), using the procedures described above, we found that teachers' rating of students' academic competence significantly and positively predicted mathematics gains, controlling for other variables. The ICC was .05, indicating that approximately 5% of the variability in students' mathematics scores was explained by the classroom they attended. For each standard deviation increase in teachers' rating, students' mathematics gains increased 3.64 W points. This translates to an effect size (*d*) of .39, which is educationally meaningful. In addition, teachers' ratings of social skills predicted an increase in mathematics score gains. We found no effect of Math PALS (i.e., no significant difference between treatment and control in mathematics) on mathematics gains; nor was there an effect of students' NSLP status or school % NSLP. As anticipated, fall word reading and mathematics scores positively predicted students' spring mathematics scores. Fall reading comprehension was not a significant predictor of mathematics gains. The results also revealed an achievement gap where students who were identified as Black or Multiracial demonstrated significantly lower scores compared with students who were identified as White. Teacher ratings of social skills were also predictive of mathematics outcomes.

Discussion

In this study, we focused on teachers' ratings of students' academic competence in the context of a post hoc data analysis of a cluster RCT, with teachers and their students randomly assigned to an intervention focusing on using assessment to guide individualized literacy instruction or to a mathematics peer-assisted learning intervention. This study elucidated several key findings. First, teachers in the A2i condition provided a more accurate rating of their students' academic competence (i.e., agreed with test scores) and were generally more effective in improving their students' literacy gains compared with teachers in the mathematics control group, but this was complicated by interaction effects. Teachers implementing Math PALS, which did not focus specifically on using assessment to inform instruction, but rather used assessment to ability group students to work in PALS dyads, generally rated the overall academic competence of their students lower compared with A2i teachers. Plus, they rated students who qualified for the NSLP as less academically competent than their more affluent peers. However, this depended on school context, with the effect (i.e., lower ratings) greater at the more affluent schools (i.e., lower percentage of students on NLSP). Second, teachers' ratings of students' academic competence predicted their students' reading and mathematics outcomes even after controlling for student, classroom, and school characteristics. Together, results revealed that teachers' ratings of their students' academic competence may influence their students' academic gains in both reading and mathematics and that teachers' perceptions may be influenced by student characteristics that are unrelated to their actual performance. Importantly, again, based on findings that control teachers tended to rate students who qualified for NSLP as less academically competent than their more affluent peers, particularly at higher SES schools, the study also reveals that we might be able to mitigate teachers' inaccurate (and sometimes negative) perceptions and potential biases about their students from higher-poverty families by providing explicit training in interpreting and using assessment results to inform instruction.

Teachers' ratings of their students' academic competence served as an important and educationally meaningful predictor of students' gains in reading and mathematics performance. This finding underscores the importance of accuracy in teachers' ratings of students' academic competence and the forging of appropriate expectations for all students. Our findings of the significant effects of SES on teacher ratings of student performance are similar to previous research (Speybroeck et al., 2012). Such implicit biases have the potential to affect teachers' behavior and consequently their interaction with students.

This may, in turn, directly affect the student-teacher relationships, the classroom learning environment, learning opportunities afforded, and, importantly, the academic and behavioral outcomes of students (Delpit, 2006; Peterson et al., 2016). Hence, building accurate teacher perceptions of students' academic competence, regardless of exogenous factors, is critical.

Our findings are inconsistent with research studies demonstrating that students' race or gender may influence teachers' perceptions of their students' academic abilities (e.g., Ready & Wright, 2011). We offer caution, however, in the interpretation of our lack of significant differences due to race; our sample largely consisted of White, non-Hispanic students, which may have limited our ability to detect significant differences in teachers' ratings of academic competence that may be due to student race/ethnicity. In addition, students who identified as Black or Multiracial tended to score lower than their White peers on mathematics outcomes but not reading outcomes. Interestingly, the A2i intervention potentially affected student outcomes, ameliorating racial achievement gaps in reading, a finding that warrants further investigation.

Consistent with previous research (Baker et al., 2015; Robinson-Cimpian et al., 2014), teachers' perception of students' social skills and behavior problems predicted teachers' ratings of academic competence. Similar to previous studies, students with stronger social skills and fewer behavior problems (as perceived by the classroom teacher) were rated as more academically capable than their peers with weaker social skills and more problem behaviors. Furthermore, we found teachers' perceptions of students' social skills to be predictive of students' mathematics (but not reading) gains over the school year. In other words, for students whose teachers participated in the A2i intervention, teachers' ratings of individual students' social skills and behavior were not predictive of their reading outcomes. Similar to the finding that student race was predictive of outcomes for mathematics but not for reading, this result highlights the potential of assessment-guided small-group instruction in helping to alleviate achievement gaps based on student characteristics. It also points to the need for future research involving teacher ratings of academic competence to further investigate teachers' perception of students' behavioral and social skills as predictive factors of academic outcomes.

In this study, the achievement tests utilized measures of specific academic abilities (e.g., reading and mathematics), whereas the teacher SSRS judgment task was less specific (e.g., judgment of students' global academic competence). According to findings from a meta-analysis on teacher judgment accuracy (Südkamp et al., 2012), incongruences in domain specificity influence the alignment between teachers' judgments and students' academic achievement. In other words, studies indicate more accurate perceptions and correlational alignment when the domain specificity of the teacher rating task and the achievement test is congruent; lower correlations have been observed when the domain specificity is incongruent. Given the post hoc nature of the present research, we were forced to use the teacher judgment metrics available. At the same time, the study findings afford a new look at the alignment of teacher ratings and test scores when using a non-domain-specific teacher judgment task and domain-specific test scores. Considering the aforementioned limitation, the present study adds value to the teacher judgment literature by using a novel approach for examining agreement between teachers' general, global perceptions of students' abilities and students' demonstrated abilities. Future research will seek to examine whether teacher judgment accuracies and subsequent impacts on students' performance vary when deliberately utilizing congruent measures of teacher judgment alongside domain-specific academic tasks and interventions.

We found a significant effect of A2i compared with Math PALS on how teachers rated their students' academic competence. A2i teachers rated their students as academically more competent overall than did Math PALS teachers. One reason might be that students in A2i classrooms made stronger gains in reading than did their peers in Math PALS classrooms, and this likely accounted for some of the differences in teachers' ratings. The other important difference between the two groups is that A2i teachers did not rate their students differently based on their NSLP status whereas the Math PALS teachers did. Both groups of teachers received the same amount and type of PD. However, the PD topics differed in focus – reading or math – and in the training on using assessment to guide individualized instruction.

We argue that it was likely the focus on interpreting and using assessment results to personalize literacy instruction that contributed to A2i teachers' more accurate ratings of their students' academic competence in both literacy and math. Early literacy and early numeracy are connected. In fact, subdomains of literacy, such as print knowledge and vocabulary, uniquely predict later numeracy scores (Purpura et al., 2011). It could be that the A2i teachers' PD afforded them a clearer picture of development in literacy subdomains, which in turn could improve their understanding of their students' numeracy abilities, to some extent.

The utility of assessment has long been the center of debate and controversy (e.g., Gilman & Reynolds, 1991; Guha et al., 2018). Our results suggest that the appropriate use of assessment can actually work to teachers' and students' benefit. Overall, teachers who used assessments to guide individualized instruction were better able to acknowledge their students' true capabilities, particularly students from higher-poverty families/schools, than were teachers who did not receive PD on how to incorporate assessment data to inform instruction. We propose that the use of valid and reliable assessment, coupled with higher expectations, may serve as a powerful and effective tool for teachers to more accurately assess the competencies and targeted learning needs of their students. Also worth noting, the tests used were the WJ reading tests, which are standardized assessments designed to diagnose learning and other disabilities. These assessments are psychometrically strong and so the teachers received more valid and reliable test results compared with results from other less valid assessments, for example, curriculum-based assessments, which are designed to provide a snapshot of students' specific skills and not their overall academic achievement. The A2i teachers were able to learn how to interpret complex assessment results with explicit PD. Test results were provided to the Math PALS teachers but they did not receive explicit PD in interpreting and using the assessment results.

In sum, our results indicate that the appropriate use of valid assessments can be beneficial for teachers to accurately evaluate their students and adjust their instructional strategies, which would in turn help students with their learning. Therefore, we recommend that educators use assessments throughout the school year to better understand their students and use assessment results to inform their instruction (i.e., personalizing instruction based on students' specific needs). Administering formative assessments as the school year progresses can provide a lot more information than summative assessments at the end of the lesson or school year about the constellations of skills each individual student brings to the learning environment.

There are limitations to this study that are worth noting. First, the analyses that were conducted for the current study were post hoc. As mentioned earlier, the primary purpose of the original study was to evaluate the efficacy of assessment-guided individualized literacy instruction, using A2i, from first through third grade. Therefore, the current research questions were not within the scope of the original RCT study. Nonetheless, the data collected were appropriate and sufficient to answer the research questions posed. Second, the sample was not nationally representative. The teacher participants (15 in the treatment condition and 13 in the control condition) were generally female, highly educated, and were not culturally diverse. Furthermore, they were experienced teachers with an average of 17 years of teaching. The student participants were also not racially or ethnically diverse, with the majority being White and non-Hispanic. This could have been why we did not find any racial or ethnic bias in contrast to previous studies on teacher perception (e.g., Schenke et al., 2017; Tani & Connor, 2018). However, results also show that making accurate assessments of students may not come naturally to teachers, including those who are more experienced. Hence, PD may be necessary to aid even experienced teachers to have accurate perceptions about their students. Moreover, because the participants of this study attended only five different schools, further examination of school-level SES and school characteristics with a larger number of schools may be necessary.

The data for the current study were collected in the 2008–2009 school year. Hence, the learning environment and classroom context from when the study was conducted may differ from those currently, which could limit the generalizability of the present study's findings. Moreover, it is noteworthy to discuss the limitations regarding the SSRS teacher questionnaire used to assess students' academic competence. As mentioned earlier, the teacher SSRS judgment task was less specific (e.g., judgment of students' global academic competence), whereas the assessments used to measure students' abilities were more specific (e.g., reading and mathematics). Moreover, the SSRS included a limited number of questions asking teachers to rate their students on their reading, math, and overall academic competency. Although we have conjectured that this questionnaire may be used as an appropriate measure to assess students' general competency, this is a limitation worth noting regarding this measure. Finally, we argue that the PD on assessment could have been the active ingredient in improving the accuracy of teachers' rating of academic competence. However, A2i PD also focused on providing reading instruction in small flexible learning groups and other aspects of effective literacy instruction. For example, it is possible that when teachers interacted with their students in small groups, they were better able to judge their capabilities. However, the argument could also be made that teachers in the Math PALS condition were free to monitor students' academic abilities while observing learning dyads similar to how A2i teachers engaged in teaching literacy in small groups. Of note, we found no effects of Math PALS on students' math outcomes. Students in both conditions had similar math scores in spring. All students in both conditions received math instruction with Saxon Math. Students in the math condition received math instruction with Math PALS along with Saxon Math. Saxon Math alone could have been more aligned with the math assessment that was used in the study rather than Saxon Math and Math PALS together. Despite these limitations, our study answers important questions that are still relevant today and provide evidence for the utility of valid and reliable student assessment data in measuring student learning and accurately guiding how teachers perceive their students' academic competence.

In summary, we found that teachers' ratings of their students' academic competence had a direct effect on students' literacy and mathematics achievement, even when controlling for previous literacy achievement, mathematics achievement, behavior problems, student SES, and other student, classroom, and school characteristics. Importantly, teachers participating in PD on data-driven personalized instruction were significantly more accurate in their judgments of their students' academic competence than were control teachers. Thus, we argue that the judicious use of assessment offers a way to improve student achievement by allowing more tailored instruction based on students' constellation of skills, and by improving the accuracy of teachers' perceptions of their students' abilities. This is particularly important for children who are typically underserved by schools—children living in poverty.

The educational landscape continues to shift. There remains continuing resistance to student assessments and overtesting. At the same time, teachers' increasing access to standardized student test scores may be beneficial as long as teachers know how to interpret and use these test scores. Given the pervasive academic achievement gaps observed in the K–12 public school system, findings outlined in this study promote the practical use of assessments to (a) focus on individual students as a means of providing informed instruction and (b) potentially reduce teachers' perceptions of achievement based on students' SES and other characteristics, such as race/ethnicity and gender. These results highlight the idea that improving the accuracy of teachers' judgments of students' academic capabilities might aid in mitigating circumstances that would otherwise hinder the egalitarian instruction that students should receive. Thus, we recommend and support a continuing narrative surrounding the ways in which students from varying backgrounds may be afforded equitable educational and learning experiences.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We dedicate this paper to the memory of our friend, mentor, colleague, and inspiration, Dr. Carol McDonald Connor.

We thank ISI lab members past and present, and the parents, teachers, and students, without whom this study would not have been possible. Funding was provided by the National Institute of Child Health and Human Development, R01HD48539 and P50 HD052120, the U.S. Department of Education, Institute of Education Sciences, R305A160404, R305H04013, R305B070074, and R305F100027, and the US Department of Education, Office of Innovation and Improvement, U411A17011. Views represented are ours and do not necessarily represent

the views of the funders. Dr. Connor had a financial interest in Learning Ovations, which University of California, Irvine monitored and deemed did not interfere with the rigor of this research.

References

Al Otaiba S, Connor CM, Folsom JS, Wanzek J, Greulich L, Schatschneider C, & Wagner RK (2014). To wait in Tier 1 or intervene immediately: A randomized experiment examining first-grade response to intervention in reading. Exceptional Children, 81(1), 11–27. 10.1177/0014402914532234 [PubMed: 25530622]

Baker CN, Tichovolsky MH, Kupersmidt JB, Voegler-Lee ME, & Arnold DH (2015). Teacher (mis)perceptions of preschoolers' academic skills: Predictors and associations with longitudinal outcomes. Journal of Educational Psychology, 107(3), 805–820. 10.1037/edu0000008 [PubMed: 26538767]

Black P. (2015). Formative assessment – an optimistic but incomplete vision. Assessment in Education: Principles, Policy & Practice, 22(1), 161–177.

Bos CS, Mather N, Narr RF, & Babur N. (1999). Interactive, collaborative professional development in early literacy instruction: Supporting the balancing act. Learning Disabilities Research & Practice, 14(4), 227–238.

Burns MK, Codding RS, Boice CH, & Lukito G. (2010). Meta-analysis of acquisition and fluency math interventions with instructional and frustration level skills: Evidence for a skill-by-treatment interaction. School Psychology Review, 39(1), 69–83.

Community-University Partnership for the Study of Children, Youth, and Families. (2011). Review of the Social Skills Rating System (SSRS). https://www.ualberta.ca/community-university-partnership/media-library/community-university-partnership/resources/tools---assessment/ssrsmay-2012.pdf

Connor CM (2013). Method for recommending a teaching plan in literacy education (US Patent No. 8,506,304). United States Patent Office.

Connor CM, Morrison FJ, Fishman BJ, Crowe EC, Al Otaiba S, & Schatschneider C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. Psychological Science, 24(8), 1408–1419. 10.1177/0956797612472204 [PubMed: 23785038]

Connor CM, Morrison FJ, Fishman BJ, & Schatschneider C. (2013). Assessment and instruction connections: The implications of child X instruction Interactions effects on student learning. In Sabatini J. & Albro ER (Eds.), Assessing Reading in the 21st Century: Aligning and Applying Advances in the Reading and Measurement Sciences. Lanham, MD: R&L Education.

Connor CM, Morrison FJ, Fishman BJ, Schatschneider C, & Underwood P. (2007). Algorithm-guided individualized reading instruction. Science, 315(5811), 464–465. 10.1126/science.1134513 [PubMed: 17255498]

Delpit L. (2006). Lessons from teachers. Journal of Teacher Education, 57(3), 220–231. 10.1177/0022487105285966

Desert M, Preaux M, & Jund R. (2009). So young and already victims of stereotype threat: Socio-economic status and performance of 6 to 9 years old children on Raven's progressive matrices. European Journal of Psychology of Education, 24(2), 207–218. 10.1007/bf03173012

Dunn KE, & Mulvenon SW (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. Practical Assessment, Research & Evaluation, 14(7), 1–11.

Eckert TL, & Arbolino LA (2005). The role of teacher perspectives in diagnostic and program evaluation decision-making. In Brown-Chidsey R. (Ed.), Assessment for Intervention: A Problem-Solving Approach (pp. 65–81). New York, NY: Guilford.

Fuchs LS, Fuchs D, Hamlett CL, Phillips NB, Karns K, & Dutka S. (1997). Enhancing students' helping behavior during peer-mediated instruction with conceptual mathematical explanations. The Elementary School Journal, 97(3), 223–249.

Fuchs LS, Fuchs D, Phillips NB, & Karns K. (n.d.). Peer-Assisted Learning Strategies (PALS) intervention manual 1st Grade. https://frg.vkcsites.org/what-is-pals/pals_math_manuals/

Gilman DA & Reynolds LL (1991). The side effects of statewide testing. Contemporary Education, 62(4), 273–278.

Graue ME, Ryan S, Nocera A, Northey K, & Wilinski B. (2017). Pulling preK into a K-12 orbit: the evolution of preK in the age of standards. Early Years: An International Research Journal, 37(1), 108–122. 10.1080/09575146.2016.1220925

Gresham FM, & Elliot SN (1990). Social Skills Rating System. Bloomington, MN: Pearson Assessments.

Guha R, Wagner T, Darling-Hammond L, Taylor T, & Curtis D. (2018). The promise of performance assessments: Innovations in high school learning and college admission. Learning Policy Institute.

Hill C, Bloome H, Black AR, & Lipsey MW (2008). Empirical benchmarks for interpreting effect sizes in research. Child Development Perspectives, 2(3), 172–177. 10.1111/j.1750-8606.2008.00061.x

Hinnant JB, O’Brien M, & Ghazarian SR (2009). The longitudinal relations of teacher expectations to achievement in the early school years. Journal of Educational Psychology, 101(3), 662–670. 10.1037/a0014306 [PubMed: 20428465]

Hunter SB (2019). New evidence concerning school accountability and mathematics instructional quality in the No Child Left Behind era. Educational Assessment, Evaluation and Accountability, 31, 409–436. 10.1007/s11092-019-09307-6

Jaffe LE (2009). Development, interpretation, and application of the W score and the relative proficiency index. Woodcock-Johnson III Assessment Service Bulletin No. 11. Riverside.

Machts N, Kaiser J, Schmidt FT, & Möller J. (2016). Accuracy of teachers’ judgments of students’ cognitive abilities: A meta-analysis. Educational Research Review, 19, 85–103. 10.1016/j.edurev.2016.06.003

Meissel K, Meyer F, Yao ES, & Rubie-Davies CM (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. Teaching and Teacher Education, 65, 48–60. 10.1016/j.tate.2017.02.021

Moon TR (2005). The role of assessment in differentiation. Theory into Practice, 44(3), 226–233. 10.1207/s15430421tip4403_7

Peterson ER, Rubie-Davies C, Osborne D, & Sibley C. (2016). Teachers’ explicit expectations and implicit prejudiced attitudes to educational achievement: Relations with student achievement and the ethnic achievement gap. Learning and Instruction, 42, 123–140. 10.1016/j.learninstruc.2016.01.010

Popham WJ (2001). Teaching to the test. Educational Leadership, 58(6), 16–20.

Purpura DJ, Hume LE, Sims DM, & Lonigan CJ (2011). Early literacy and early numeracy: The value of including early literacy skills in the prediction of numeracy development. Journal of Experimental Child Psychology, 110(4), 647–658. [PubMed: 21831396]

Raudenbush SW, & Bryk AS (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Sage.

Ready DD, & Wright DL (2011). Accuracy and inaccuracy in teachers’ perceptions of young children’s cognitive abilities: The role of child background and classroom context. American Educational Research Journal, 48(2), 335–360. 10.3102/0002831210374874

Robinson-Cimpian JP, Lubienski ST, Ganley CM, & Copur-Gencturk Y. (2014). Teachers’ perceptions of students’ mathematics proficiency may exacerbate early gender gaps in achievement. Developmental Psychology, 50(4), 1262–1281. 10.1037/a0035073 [PubMed: 24294875]

Roehrig AD, Duggar SW, Moats LC, Glover M, & Mincey B. (2008). When teachers work to use progress monitoring data to inform literacy instruction: Identifying potential supports and challenges. Remedial and Special Education, 29, 364–382. 10.1177/0741932507314021

Rosenthal R, & Jacobson L. (1968). Pygmalion in the classroom: Teacher expectations and pupils’ intellectual development. Holt, Rinehart & Winston.

Rubie-Davies CM (2010). Teacher expectations and perceptions of student attributes: Is there a relationship? British Journal of Educational Psychology, 80(1), 121–135. 10.1348/000709909X466334

Rubie-Davies CM, Peterson ER, Sibley CG, & Rosenthal R. (2015). A teacher expectation intervention: Modelling the practices of high expectation teachers. Contemporary Educational Psychology, 40, 72–85.

Schenke K, Nguyen T, Watts TW, Sarama J, & Clements DH (2017). Differential effects of the classroom on African American and non-African American's mathematics achievement. Journal of Educational Psychology, 109(6), 794–811. 10.1037/edu0000165 [PubMed: 28824200]

Speybroeck S, Kuppens S, Van Damme J, Van Petegem P, Lamote C, Boonen T, & de Bilde J. (2012). The role of teachers' expectations in the association between children's SES and performance in kindergarten: A moderated mediation analysis. PLoS One, 7(4), e34502. 10.1371/journal.pone.0034502 [PubMed: 22506023]

Südkamp A, Kaiser J, & Möller J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. Journal of Educational Psychology, 104, 743–762. 10.1037/a0027627

Südkamp A, Praetorius AK, & Spinath P. (2017). Teachers' judgment accuracy concerning consistent and inconsistent student profiles. Teaching and Teacher Education, 76, 204–213. 10.1016/j.tate.2017.09.016

Tani N, & Connor CM (2018). Teachers' judgments of 2nd graders' academic competence and impacts on reading comprehension: Racial and gender differences [Poster presentation]. American Educational Research Association Annual Meeting, New York.

Timmermans AC, de Boer H, & van der Werf MPC (2016). An investigation of the relationship between teachers' expectations and teachers' perceptions of student attributes. Social Psychology of Education, 19(2), 217–240. 10.1007/s11218-015-9326-6

Wininger RS (2005). Using your tests to teach: Formative summative assessment. Teaching Psychology, 32(2), 164–166.

Woodcock RW, McGrew KS, & Mather N. (2001). Woodcock-Johnson-III tests of achievement. Riverside.

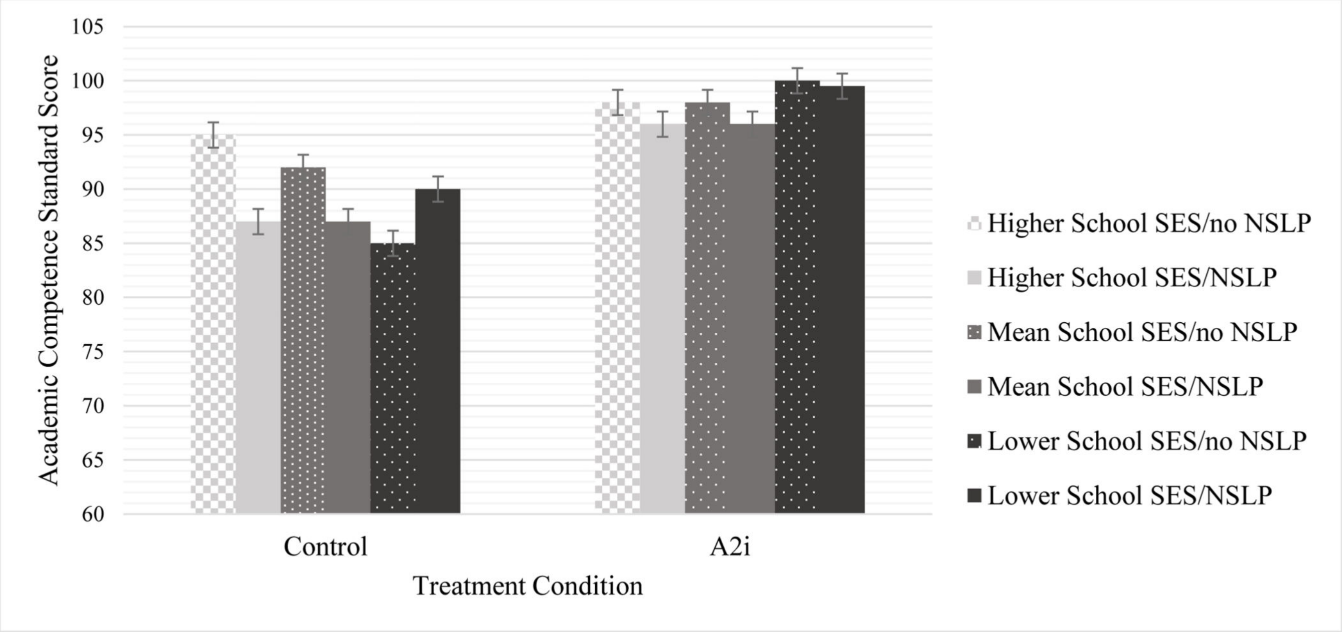

Figure 1. HLM modeled results of teachers' fitted mean rating of students' academic competence.

Note. This figure demonstrates the HLM modeled results of teachers' fitted mean rating of students' academic competence as a function of treatment condition (Control & A2i), Percent of students at the school who qualified for the National School Lunch Program (school NSLP %) modeled at the 25th (more affluent, light gray), 50th (medium gray) and 75th (dark gray – more disadvantaged) of the sample where the mean school NSLP % was 37%, and student eligibility for the National School Lunch Program (NSLP, yes or no). Error bars are standard errors.

Table 1

Sample Means and Standard Deviations Totals and by Treatment Condition

		Mean	Std. Deviation	Std. Error	Minimum	Maximum
Fall Letter-Word Identification	Math PALS	417.78	27.361	1.954	367	509
	A2i	422.19	27.140	1.745	369	503
	Total	420.21	27.296	1.304	367	509
Fall Passage Comprehension	Math PALS	441.34	22.421	1.602	358	491
	A2i	444.17	22.956	1.476	377	488
	Total	442.90	22.736	1.086	358	491
Fall Applied Problems	Math PALS	445.95	14.714	1.051	405	498
	A2i	446.90	13.861	.891	405	485
	Total	446.47	14.240	.680	405	498
Spring Letter-Word Identification	Math PALS	457.91	22.526	1.609	408	531
	A2i	466.11	20.079	1.291	400	512
	Total	462.44	21.574	1.031	400	531
Spring Passage Comprehension	Math PALS	472.02	13.519	.966	438	508
	A2i	477.46	12.585	.809	425	505
	Total	475.02	13.276	.634	425	508
Spring Applied Problems	Math PALS	464.76	16.464	1.176	428	512
	A2i	464.74	15.228	.979	424	498
	Total	464.75	15.774	.754	424	512
Teacher Ratings of Academic Competence SS	Math PALS	92.46	22.780	1.627	0	115
	A2i	98.17	13.217	.850	60	115
	Total	95.62	18.330	.876	0	115
Teacher Rating of Social Skills SS	Math PALS	100.59	19.685	1.406	0	130
	A2i	101.30	19.370	1.245	0	130
	Total	100.98	19.492	.931	0	130
Teacher Rating of Problem Behavior SS	.00	103.26	15.292	1.092	85	142
	1.00	99.76	14.217	.914	85	141
	Total	101.32	14.793	.707	85	142

Note. W = W score; SS = standard score.

Table 2

Correlations among Teacher Ratings and Standardized Achievement Assessments

	1	2	3	4	5	6	7
1. SSRS Academic Competence		**.621**	**.651**	**.440**	**.731**	**.669**	**.581**
2. Fall Letter Word Identification	.428		**.797**	**.449**	**.776**	**.636**	**.551**
3. Fall Passage Comprehension	.415	.797		**.424**	**.689**	**.658**	**.473**
4. Fall Applied Problems	.323	.449	.424		**.432**	**.422**	**.683**
5. Spring Letter Word Identification	.492	.776	.689	.453		**.784**	**.509**
6. Spring Passage Comprehension	.331	.636	.658	.422	.784		**.494**
7. Spring Applied Problems	.455	.551	.473	.683	.509	.494	

Note. All correlations are significantly greater than 0 with $p < .001$. Bold and above the diagonal are correlations for the A2i treatment group; below the diagonal are correlations for the control group. SSRS = Social Skills Rating System.

Table 3

HLM Analysis Predicting Teachers' Ratings of Students' Academic Competence

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
Fitted Mean Academic Competence	91.906	4.393	20.918	24	<0.001
A2i Classroom	5.142	5.957	0.863	24	0.397
School Percent NSLP	−1.008	.794	−1.268	24	0.217
A2i X School Percent NSLP	1.241	1.092	1.137	24	0.267
Girl	−.287	.952	−.302	399	0.763
NSLP	−1.882	1.850	−1.017	399	0.310
NSLP X A2i Classroom	1.287	2.288	.562	399	0.574
NSLP X School Percent NSLP	1.263	.358	3.528	399	<0.001
NSLP X A2i X School Percent NSLP	−1.286	.441	−2.917	399	0.004
Fall Word Reading	.335	.035	9.464	399	<0.001
Fall Reading Comprehension	.102	.033	3.059	399	0.002
Social Skills Rating	.078	.036	2.132	399	0.034
Behavior Problems Rating	−.228	.049	−4.605	399	<0.001
Word Reading Gains	.285	.038	7.498	399	<0.001
Math Gains	.090	.040	2.253	399	0.025
Random Effect	**Standard Deviation**	**Variance Component**	***d.f.***	χ^2	***p*-value**
Classroom	15.330	235.026	24	566.908	<0.001
Student	9.426	88.867			

Note. Deviance = 3307.770; Intraclass correlation from unconditional model (ICC) = .520.

NSLP = National School Lunch Program students; Girl (girl = 1; boy = 0); Fall Word Reading was assessed with WJ Letter Word Identification subtest of the Woodcock Johnson III Test of Achievement (WJ); Fall Reading Comprehension was assessed with Passage Comprehension subtest of the WJ; Math was assessed with Applied Problems Subtest of the WJ; Academic Competence, Social Skills, and Behavior Problems were measured with Social Skills Rating System and standard scores were used in the analysis. Word Reading Gains and Math Gains are the difference in W score between spring and fall scores.

Table 4

Predicting Students' Reading Comprehension Outcomes in the Spring

Fixed Effect	**Coefficient**	**Standard error**	***t*-ratio**	**Approx. *d.f.***	***p*-value**
Fitted Mean Reading Comp.	472.130	0.939	502.640	25	<0.001
A2i Classroom	4.435	1.036	4.281	25	<0.001
School Percent NSLP	0.055	0.096	0.573	25	0.572
Girl	1.290	0.927	1.392	398	0.165
NSLP	1.386	1.028	1.348	398	0.178
Fall Word Reading	0.096	0.028	3.415	398	<0.001
Fall Reading Comprehension	0.138	0.033	4.167	398	<0.001
Fall Math	0.088	0.037	2.368	398	0.018
Social Skills Rating	0.060	0.037	1.644	398	0.101
Academic Competence Rating	0.238	0.044	5.373	398	<0.001
Behavior Problems Rating	0.017	0.050	0.334	398	0.738
Asian	−1.125	3.208	−0.351	398	0.726
Black	−0.119	2.141	−0.056	398	0.956
Hispanic	−3.415	2.670	−1.279	398	0.202
Multiracial	−1.877	1.960	−0.957	398	0.339
Random Effect	**Standard Deviation**	**Variance Component**	***d.f.***	χ^2	***p*-value**
Classroom	1.247	1.556	25	31.434	0.175
Student	9.293	86.359			

Note. Deviance = 3198.748; Intraclass correlation from unconditional model (ICC) = .052.

NSLP = National School Lunch Program; A2i classroom (A2i = 1; control = 0); Girl (girl = 1, boy = 0); Student NSLP status (Qualified for NSLP = 1); White is the fixed reference group for race/ethnicity. Fall Word Reading was assessed with WJ Letter Word Identification subtest of the Woodcock Johnson III Test of Achievement (WJ); Fall Reading Comprehension was assessed with Passage Comprehension subtest of the WJ; Math was assessed with Applied Problems Subtest of the WJ; Academic Competence, Social Skills, and Behavior Problems were measured with Social Skills Rating System and standard scores were used in the analysis. Word Reading Gains and Math Gains are the difference in W score between spring and fall scores.

Table 5

Predicting Students Mathematics Outcomes in the Spring

Fixed Effect	**Coefficient**	**Standard error**	***t*-ratio**	**Approx. *d.f.***	***p*-value**
Fitted Mean Spring Math	466.817	1.328	351.457	25	<0.001
A2i Classroom	−0.846	1.607	−0.526	25	0.603
School Percent NSLP	−0.281	0.148	−1.897	25	0.069
Gender (Boy = 0)	−1.832	1.022	−1.792	398	0.074
NSLP	1.404	1.144	1.228	398	0.220
Fall Word Reading	0.073	0.031	2.305	398	0.022
Fall Reading Comprehension	0.013	0.036	0.350	398	0.726
Fall Math	0.556	0.041	13.621	398	<0.001
Social Skills Rating	0.092	0.040	2.300	398	0.022
Academic Competence Rating	0.198	0.049	4.037	398	<0.001
Behavior Problems Rating	0.033	0.055	0.602	398	0.547
Asian	−0.116	3.546	−0.033	398	0.974
Black	−5.875	2.383	−2.466	398	0.014
Hispanic	0.620	2.949	0.210	398	0.833
Multiracial	−5.761	2.179	−2.644	398	0.009
Random Effect	**Standard Deviation**	**Variance Component**	***d.f.***	χ^2	***p*-value**
Classroom	3.253	10.584	25	63.283	<0.001
Student	10.169	103.407			

Note. Deviance = 3292.090; Intraclass correlation from unconditional model (ICC) = .05.

NSLP = National School Lunch Program, A2i classroom (A2i=1; control =0), Girl (girl = 1, boy = 0), Student NSLP status (NSLP = 1); White is the fixed reference group. Fall Word Reading was assessed with WJ Letter Word Identification subtest of the Woodcock Johnson III Test of Achievement (WJ); Fall Reading Comprehension was assessed with Passage Comprehension subtest of the WJ; Math was assessed with Applied Problems Subtest of the WJ; Academic Competence, Social Skills, and Behavior Problems were measured with Social Skills Rating System and standard scores were used in the analysis. Word Reading Gains and Math Gains are the difference in W score between spring and fall scores.