2017

Measuring teacher ability to predict student success in solving specific mathematics problems: Procedures and initial findings of accuracy, overprediction, and underprediction

Robert C Schoen and Naomi Iuhasz-Velez

Measuring Teacher Ability to Predict Student Success in Solving Specific Mathematics Problems

Procedures and Initial Findings of Accuracy, Overprediction, and Underprediction

Robert C. Schoen
Naomi Iuhasz-Velez

MAY 2017

Research Report No. 2017-03

The research and development reported here were supported by the Institute of Education Sciences, U.S. Department of Education, through Award No. R305A120781 to Florida State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Suggested citation: Schoen, R. C., & Iuhasz-Velez, N. (2017). *Measuring teacher ability to predict student success in solving specific mathematics problems: Procedures and initial findings of accuracy, overprediction, and underprediction.* (Research Report No. 2017-03). Tallahassee, FL: Learning Systems Institute, Florida State University. 10.17125/fsu.1507903318

Measuring Teacher Ability to Predict Student Success in Solving Specific Mathematics Problems

Procedures and Initial Findings of Accuracy, Overprediction, and Underprediction

Research Report No. 2017-03

Robert C. Schoen

Naomi Iuhasz-Velez

May 2017

Florida Center for Research in Science, Technology, Engineering, and Mathematics (FCR-STEM)
Learning Systems Institute
Florida State University
Tallahassee, FL 32306

Acknowledgements

In addition to the important support of the Institute of Education Sciences, the successful collection of the data needed for the present study involved many, many people. Some of the most critical are listed below, along with their roles.

Robert Schoen was integrally involved with designing and implementing the interview and interviewer training, selecting the items for the teacher judgment accuracy instrumentation, report writing, and overall management of the various components of the larger study. Naomi Iuhasz-Velez conducted much of the literature review and data analysis for this report and contributed heavily to the writing and interpretation of results.

Amanda Tazaz coordinated the on-site data collection each spring and deserves tremendous credit for finding teachers and students in schools on the data-collection days. She assisted Robert Schoen in making decisions about which items to include on the teacher judgment accuracy instrument. She also provided important knowledge about the actual events on data-collection days (e.g., explaining how we came to have data corresponding to six students in one classroom rather than the expected four).

Walter Secada, Juli Dixon, Mark LaVenia, and Kristopher Childs were integrally involved in weekly management-team meetings throughout the startup and implementation of the larger study.

Zachary Champagne worked closely with Robert Schoen on the design of the 2015 MPAC interview, training of interviewers in spring 2015, and selection of items for the teacher judgment accuracy instrument in spring 2015.

Along with Amanda Tazaz, Kristopher Childs coordinated the interview team and data collection for both teachers and students in 2014 and 2015. Other members of the interview team included Charity Bauduin, Wendy Bray, Anne Brown, Zachary Champagne, Rebecca Gault, Vernita Glenn-White, Katie Harshman, Karon Kerr, Edward Knote, Erika Moore, Magnolia Placido, Nesrin Sahin, Melissa Soto, Makini Sutherland, Laura Tapp, Harlan Thrailkill, Gillian Trombley, Alex Utecht, Pooja Vaswani, and Ian Whitacre. Video coding of the interviews was conducted by Robert Schoen, Zachary Champagne, Kristy Farina, Shelby McCrackin, Ian Whitacre, and Nesrin Sahin.

In addition to managing the data, Kristy Farina created data-entry systems, oversaw verification of data accuracy and security, curated the data, and helped with data analysis for the present report. She also assisted with data requests from schools and districts throughout the study and assisted with report writing.

Anne Thistle provided valuable assistance with editing, and Casey Yu managed the final layout, formatting, and style.

We are especially grateful to the Institute of Education Sciences at the U.S. Department of Education for their support and to the students, parents, principals, district leaders, and teachers who agreed to participate in the study and contribute to advancing knowledge in mathematics education. Without them, this work would not be possible.

Table of Contents

List of Appendices

List of Tables

List of Figures

List of Equations

Executive Summary

A teacher's ability to predict his or her students' performance is referred to as *teacher judgment accuracy*. We conceptualize teacher judgment accuracy to be a type of teacher knowledge. This report describes efforts to measure teachers' knowledge of their own students' abilities in mathematics.

The purpose of this report is to describe the data-collection procedures we used in our attempts to measure teacher judgment accuracy. We provide information about the items used, the data-collection procedure, and initial analyses of the results using descriptive statistics based on classical test theory. We anticipate that teacher judgment accuracy will ultimately yield important insight into the elusive link between teacher knowledge and student learning.

For brevity, we refer to the assessment instrument described in this report as the Teacher Ability to Predict Student Success (TAPSS) instrument. We anticipate that data generated with the TAPSS instrument will be used to determine the effects of a teacher professional-development program on teacher judgment accuracy (and the related ideas of incorrect prediction a student would produce a correct or incorrect answer, which has been called overprediction or underprediction), and the extent to which this type of knowledge is associated with student learning.

Context

The work described here was completed as part of a randomized controlled trial evaluating the implementation and impact of a teacher professional-development program called Cognitively Guided Instruction (CGI). The present report provides a description of the sample, a description of the study design and its realization, and descriptive statistics. These statistics summarize the data to highlight emerging patterns concerning teacher judgment accuracy, overprediction, and underprediction both overall and on individual items.

Student data for the TAPSS instrument was gathered through one-on-one mathematics interviews conducted in spring 2014 and spring 2015 (Schoen, LaVenia, Champagne, & Farina, 2016; Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016). During each wave of data collection, teachers were shown a set of problems from the interview and were told which students would be attempting to solve the problems. For each student and item, the teachers predicted whether the student would solve the item correctly. The analytic sample for the spring 2014 wave of data collection involved grade 1 and grade 2 students (n = 504) and their mathematics teachers (n = 146). The analytic sample for the 2015 wave of data collection included a larger sample of students (n = 785) and their teachers (n = 200). The measurement of teachers' knowledge of their students was conducted at the end of that school year, and the students had been in their corresponding teachers' classrooms for a full school year.

Advancements in the Measurement of Teacher Judgment Accuracy

The TAPSS instrument and the sample described here represent several advancements in the field of measuring teacher judgment accuracy. Compared to other instruments measuring a similar construct (cf. Carpenter, Fennema, Peterson, Chiang, & Loef, 1989; Gabriele, Joram, & Park, 2016), the instrument used in the present study includes fewer items and presents a smaller burden on the test takers. The TAPSS instrument is more specific than other extant measures in three major ways. First, the teachers predicted the success of individual students rather than that of their whole class. Second, they predicted those students' success on individual items, rather than predicting an overall test score. Third, accuracy and inaccuracy were analyzed by comparison of teachers' predictions with students' answers for each of the individual items.

Instrument Development

The TAPSS instrument was developed over a period of two years involving two large-scale field tests. The present report provides information about the items used, the data-collection procedure, and initial descriptive statistics. In the first year (spring 2014), teachers predicted whether individual students would correctly solve each of four computation items and also the students' cognitive process, i.e. whether they would know the basic addition and subtraction facts at a recall level. Among the spring 2015 item set were word problems, problems involving basic addition and subtraction facts, problems involving multidigit computation, and problems exploring student understanding of the meaning of the equals sign. In the second year (spring 2015), teachers predicted whether students would solve each of seven items correctly, but they were not asked to predict whether students knew the basic facts at a recall level. This change was based on concerns that teachers did not share a common understanding of the phrase “recalled facts” and so as to focus and clarify the construct of interest.

Findings and Next Steps

Development of the TAPSS instrument benefited from the opportunity to administer it twice. More than 80% of the students solved three of the four items in the spring 2014 TAPSS instrument correctly, and the teachers accurately predicted the performance of almost every student on almost every item. We therefore, in addition to discontinuing teacher prediction of cognitive processes, revised the instrument in 2015 to increase the number of items and to include more difficult items. Across all items and both grade levels, students in the 2015 sample correctly solved the problems 51% of the time.

Teachers in the sample were highly accurate in their predictions of student performance on the items involving basic number facts, but they were much less accurate on the items related to multidigit computation, the meaning of the equals sign, and word problems.

The 2014 sample of teachers evidenced very little variation in their accuracy, overprediction, and underprediction data. The 2015 sample of teachers showed much greater variation. This result may be due, in part, simply to the greater number of items in 2015 (n = 7) than in 2014 (n = 4). It is also due to a careful selection of items that (a) were more difficult for students, (b) covered a range of topics in the mathematics curriculum, and (c) represented topics in which teachers are often surprised to learn what their students do or do not know (e.g., meaning of the equals sign).

The overall rate of accurate predictions was higher in the grade 2 sample than in the grade 1 sample, probably because higher percentages of grade 2 students solved the items correctly (six of the seven items were identical to those in grade 1). Future work in this area may be well advised to include some items at grade 2 that are slightly more difficult for students to solve correctly (i.e., would result in a lower percentage of correct solutions by grade 2 students).

The development of the TAPSS instrument occurred in the context of a larger study. The data for the larger study includes information about these same individual students’ characteristics, such as performance on standardized tests, gender, race and ethnicity, and exceptionality. It also includes information about teacher knowledge and beliefs, years of experience, educational background, and more. The size and scope of the set of data create many options for research on the topic of teacher judgment accuracy based on the existing data. More such research is forthcoming.

1. Introduction and Overview

For many years now, researchers have sought to identify and measure facets of knowledge that may predict or explain differences in teacher effectiveness (Ball, Thames, & Phelps, 2008; Campbell et al., 2014; Hill, Rowan, & Ball, 2005; Schoen, Bray, Wolfe, Tazaz, & Nielsen, 2017; Shulman, 1986). Almost all of the existing constructs and instruments designed to measure this knowledge focus on generalized knowledge of subject matter and students. Relatively little research has focused on knowledge relevant to individual student differences on a specific mathematics problem at a specific point in time.

The Cognitively Guided Instruction (CGI) program focuses teachers' attention on details in their own students' cognitive processes with respect to a generalized framework for progression in children's mathematical thinking (Carpenter et al., 1989; Carpenter & Franke, 2004; Fennema et al., 1996). As a component of an efficacy study evaluating variation in implementation and impact of a CGI teacher professional-development program, we gathered data on teachers' knowledge of their own students' abilities to solve specific mathematics problems on a specific day.

Recently published studies have developed a related construct called *teacher judgment accuracy* (Gabriele et al., 2016; Südkamp, Kaiser, & Möller, 2012). A previous randomized controlled trial of a CGI program attempted to measure this type of teacher knowledge (Carpenter et al., 1989). The present study, named *Replicating the CGI Experiment in Diverse Environments,* used an adapted procedure in an attempt to measure the same construct: teachers' knowledge of the mathematical thinking of individual students in their classes.

The purpose of the present report is to describe the data-collection procedures for measuring teacher judgment accuracy as part of the *Replicating the CGI Experiment in Diverse Environments* study. The CGI program is intended to increase teacher knowledge of their own students' cognitive processes as a critical step in guiding instructional decisions. Teacher judgment accuracy was measured and defined to be an outcome of interest in the evaluation of the CGI program. Because teachers' knowledge of their own students is viewed as an important component in teaching, teacher judgment accuracy could also be considered a mediator of the effect of the intervention program on student achievement.

1.1. Rationale

Teacher judgment accuracy is defined as the extent to which teachers' predictive perceptions of children's academic skills are consistent with objective assessments of their students' skills (Gabriele et al., 2016; Ready & Wright, 2011; Südkamp et al., 2012). We conceptualized teacher judgment accuracy as a component of teacher knowledge that might yield meaningful and useful insight into the link between teacher knowledge and student learning.

The construct of teacher judgment accuracy differs from general knowledge of typical learning progressions (Carpenter, Fennema, Franke, Levi, & Empson, 2015; Sarama & Clements, 2009). Essentially, the distinction is between generalizable knowledge of typical students on the one hand and specific (and timely) knowledge of individual students on the other. In the act of teaching, teachers' knowledge of their individual students probably does not replace generalizable knowledge of students and subject matter, but it may supplement it and increase the effect of generalizable knowledge on teacher effectiveness.

Teachers' accuracy in their assessment of their students' knowledge and abilities is important to day-to-day instructional decisions. We therefore view teacher judgment accuracy to be an integral component of the formative assessment process. This type of knowledge of individual students may be an important

mediating factor with respect to the effect of formative assessment on student learning (Lang, Schoen, LaVenia, & Oberlin, 2014; Lang, Schoen, LaVenia, Oberlin, & Robinson, 2013). We conjecture that higher levels of teacher judgment accuracy might result in instruction that is better aligned to student needs. Conversely, teachers' overestimation or underestimation of student knowledge and abilities may create a mismatch between instruction and student knowledge that may result in suboptimal learning opportunities for students.

High teacher expectations for student knowledge attainment are generally considered to be desirable, but unrealistically high expectations for students' abilities in the moment may have detrimental effects. If a teacher infers or assumes that students have much higher knowledge or ability than they actually do in the moment of teaching, the teacher may be offering instruction that is too advanced to be useful in the learning process. If a teacher infers or assumes that students do not understand mathematical concepts that the students already do understand, the teacher may waste valuable instructional time on concepts or skills that do not help to advance student abilities. Both of these scenarios may initiate a cascade of undesirable effects on the learner. In the measurement of teachers' estimation of their students' abilities, we refer to teachers' overestimation of their students' abilities as *overprediction* and to underestimation of their students' abilities as *underprediction*.

1.2. Brief Overview of Research on Teacher Estimation of Student Abilities

In the first systematic examination of teacher judgment accuracy in research, Hoge and Coladarci (1989) reviewed 16 published studies regarding the relation between teachers' predictions of student performance and students' actual performance on a concurrent, independent criterion of achievement. They reported a moderately high correlation (.66) between teachers' judgments and students' achievement on a standardized test. A more recent metaanalysis of teacher judgments (Südkamp et al., 2012) found a similar positive and high correlation between teachers' predictions of students' performance and students' actual performance on standardized achievement tests. The average correlation coefficient was r = .63 over 75 studies. Although these studies further investigated theoretically and methodologically relevant moderators of this correlation, efforts to measure teacher judgment accuracy have not yet resulted in consensus on the superiority of any specific method for measuring the construct (Gabriele et al., 2016; Südkamp et al., 2012).

A large corpus of research on teachers' knowledge of their students has largely focused on deficits in teacher knowledge rather than on accuracies in teacher estimation. Researchers focused on teachers' inaccuracy, bias, and its broader social implications have come to call it the *Pygmalion effect* or *self-fulfilling prophecy* (de Boer, Bosker, & van der Werf, 2010; Jussim & Eccles, 1992; Madon, Jussim, & Eccles, 1997). The focus on erroneous or inaccurate teacher knowledge of student abilities can be traced to a seminal study by Rosenthal and Jacobson (1968) published in the book titled *Pygmalion in the Classroom*. Rosenthal and Jacobson found that experimentally induced, incorrect teacher expectations affected students' IQ scores. They found these expectation effects to still be present and measureable two years later. Expectation effects are hypothesized to occur when a teacher's perception of a student's motivation or ability level differs from the student's self-reported motivation and ability level. If this discrepancy is found to be related to the student's subsequent achievement (when prior achievement is controlled for), the teacher's bias is assumed to have become a self-fulfilling prophecy (Harvey, Suizzo, & Jackson, 2016).

Very little extant research on this topic focuses on teacher knowledge and accuracy. In our assessment, the corpus of research literature related to bias, prejudice, and expectations largely ignores teachers' accurate knowledge of their individual students. The focus on teachers' biased expectations suggests a deficit-based interpretation of teacher judgment, where teacher misjudgment may suppress student

learning. On the contrary, a focus on teacher accuracy suggests an asset-based interpretation, where teacher judgment may have a positive effect on student learning. Hoge and Coladarci (1989) credit Carpenter et al. (1989) for being among the first to link individual differences among teachers' knowledge of their own students to differences in teacher effectiveness. In contrast with the deficit model that dominates the literature on teacher knowledge of their students' abilities, we note that Carpenter et al. focused on teachers' correct knowledge of their students, which is consistent with their recommended approach to assessing students' cognitive processes (Carpenter et al., 2015).

Because of the vast range of research methods observed in studies focused on teacher perceptions of their students, Südkamp et al. (2012) commented that not all studies of teacher judgment accuracy involved comparable situations. They pointed out that inaccuracy of teacher judgments may be grounded in the studies' methodologies rather than in the teachers' diagnostic competency. Through their metaanalysis, they identified several judgment characteristics of studies that moderated the degree of accuracy. They found that teachers have higher judgment accuracy for *informed judgments*—when teachers were made aware of the content of the test or the standard of comparison on which their judgment is based—than for *uninformed judgments*—wherein teachers were not apprised of the content of the test. Only 10 out of the 75 studies identified in the 2012 review qualified as *informed judgment* situations. Südkamp et al. found significantly higher correlations between teachers' predictions and students' test performance for informed (mean effect size = .76) than for uninformed judgments (mean effect size = .61).

Another factor influencing teacher judgment accuracy is *judgment specificity*. Hoge and Coladarci (1989) identified five categories, ranging from low to high specificity: rating students' academic achievement on a rating scale (e.g., poor–excellent), ranking the students in a class in order of academic achievement, assigning grade equivalence for students' performance on a standardized achievement test, estimating the number of correct responses achieved by a student on a standardized achievement test, and fifth and most specific, indicating students' item responses on each item of an achievement test. None of the 75 studies in the 2012 metaanalysis asked teachers to indicate students' responses on each item. In fact, 59 of the 70 applicable studies in their sample used the rating scale, which is the least specific.

Methodological details of the methods chosen for measuring teacher knowledge of their students' abilities can significantly affect the interpretation of the score and the relation of the score to other factors. The TAPSS assessment—the topic of the present report—involves an informed judgment situation (Südkamp et al., 2012) and conforms to the highest rating on the specificity scale created by Hoge and Coladarci (1989). The TAPSS assessment is more *specific* than most of the extant measures in this field. The measurement strategy in the TAPSS assessment is both task-specific and student-specific. Teachers were apprised of the individual problems students were asked to solve and the individual students who would solve them, making TAPSS an informed-judgment situation. Moreover, the present instrument is highly *congruent* (i.e., teachers predict item-level student performance, student performance is analyzed at the same level), a similarly unusual feature in this corpus of research (Südkamp et al., 2012). Finally, the TAPSS assessment is designed to measure both teacher accuracy and inaccuracy in their estimation of student abilities. A detailed description of the data-collection procedures, sample and setting, and some initial descriptive results based on field tests of the TAPSS assessment is provided in the following sections.

2. Methods

2.1. Description of the Sample and Setting

Data for the present study were drawn from a study involving students and teachers in 22 schools situated in two diverse public school districts in Florida. Grade 1 and grade 2 teachers in these schools voluntarily consented to participate in a large-scale, cluster-randomized controlled trial evaluating the efficacy of a teacher professional-development program in mathematics. The study period lasted two years. Recruitment of teachers and schools began in January 2013. Teacher professional development workshops started in summer 2013 and continued through the 2013–14 school year, summer 2014, and the 2014–15 school year.

The Common Core State Standards for Mathematics (NGACBP & CCSSO, 2010) drove the accountability systems in Florida at the outset of the study. The State Board of Education in Florida adopted a revised set of standards called the Mathematics Florida Standards (http://www.cpalms.org/) in the middle of the two-year study period. No statewide mathematics assessment for grade 1 or grade 2 students was conducted during the study period.

2.1.1. Randomization of Assignment of Schools to Treatment Condition

The 22 schools were blocked on district and stratified into matched pairs on the basis of the percentages of students in each school who were eligible for free or reduced-price lunch in the 2012–13 school year. One-half of the schools in the sample were assigned at random to the treatment condition; the other half were assigned to the control condition. More information about the study design, including the randomization procedures and results of random assignment, can be found in a separate report (Schoen, LaVenia, Tazaz, & Farina, 2017).

2.1.2. The Teacher Sample

Teacher judgment accuracy data were collected for 146 of the participating teachers during the first year of the study and for 200 teachers participating during the second year. Any person teaching mathematics to grade 1 or grade 2 students in the 22 participating schools was eligible to participate in the study. On average, each school had nine participating classroom teachers in the study, but the number of teachers representing each school in our sample ranged from two to 20 teachers. All of the teachers participated voluntarily; their participation was not mandated by their schools or districts.

About half of the participating classroom teachers taught first grade (53%) and the other half second grade (47%) during the study period. The teachers in the sample were predominantly female (99% during year one; 93% during year two). About three-fourths of the teachers in the sample reported having four or more years of teaching experience when they started participating in the study (81% teachers in year one; 72% in year two).

Table 22 in Appendix A reports demographic statistics for the spring 2014 teacher sample in the present study. In the present report, the data include the predictions of participating grade 1 or grade 2 teachers who completed prediction forms for students who were interviewed. These individual teachers comprise what we call the *predicting analytic sample* (*n* = 146). The table displays these statistics separated by grade and treatment-condition. Only teachers who had at least four students with positive

parental consent for video recording were asked to predict student performance in the spring 2014 wave of data collection.[1]

The teacher sample changed in size as well as other characteristics between spring 2014 and spring 2015. Some teachers left the participating schools, and some late-joiners decided to participate beginning in the second year. These changes occurred in schools in both the treatment and the control conditions. Table 23 in Appendix A reports the demographic data for the predicting analytic sample of teachers (*n* = 200) contributing data to the spring 2015 wave of data collection.

2.1.3. The Student Sample

Students in the 22 sample schools for whom parental consent was obtained completed four mathematics tests as part of their participation in the study: a whole-group-administered, written test in the fall (Schoen, LaVenia, Bauduin, & Farina, 2016a, 2016b); the Iowa Test of Basic Skills Math Problems and Math Computation tests (ITBS; Dunbar et al., 2008) administered in a whole-class setting at the end of the 2014–15 school year; and the Mathematics Performance and Cognition interview (MPAC; Schoen, LaVenia, Champagne, & Farina, 2016; Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016), which was administered in an individual, one-on-one setting at the end of the school year. Information about these tests is provided in the Measures section (Section 2.2) of the present report. Students who completed the MPAC interviews in spring 2014 or spring 2015 and for whom teachers predicted their answers for the focal items on the MPAC interviews comprise the analytic sample for the present report. To be eligible to complete the MPAC interview and, by extension, to be included in the TAPSS data set, students had to be part of the sample at the time of the fall test of that respective school year.

The students who completed the 2014 MPAC interview (Schoen, LaVenia, Champagne, & Farina, 2016) were selected through a stratified-random-sampling procedure from a larger sample composed of 2,373 students (1,226 grade 1 and 1,147 grade 2) for whom signed parental consent was obtained and who completed the fall 2013 mathematics test. The students who completed the 2015 MPAC interview were selected through a stratified-random-sampling procedure from a larger sample composed of 3,083 students (1,597 grade 1 and 1,486 grade 2) for whom signed parental consent was obtained and who completed the fall 2014 mathematics test.

Tables 1 and 2 report the sample sizes for each of the measurement instruments in the 2013–14 and 2014–15 school years, respectively. The aim of random selection was to include four students (two boys and two girls from each participating teacher's class). The class was split into two groups corresponding to the upper- and lower-performing halves of the class on the basis of the fall mathematics test. One boy and one girl were selected at random from each of the two groups within each class.

Table 1. 2013–14 Number of Students in Analytic Sample for Each Measurement Instrument

Measure	Number of students in analytic sample		
	Grade 1	Grade 2	Total
Fall 2013 mathematics test	1,226	1,147	2,373
MPAC interview	336	286	622
Teacher Judgment Accuracy	277	227	504

[1]This was not a requirement in the 2015 wave of data collection.

Table 2. 2014–15 Number of Students in Analytic Sample for Each Measurement Instrument

Measure	Number of students in analytic sample		
	Grade 1	Grade 2	Total
Fall 2014 mathematics test	1,597	1,486	3,083
MPAC interview	440	416	856
Teacher Judgment Accuracy	420	367	785

Tables 24 and 25 in Appendix A report the demographics for the sample of participating students who completed the MPAC interviews and for whom their teachers predicted their answers for the 2014 and 2015 target items. The samples in the tables were also split by treatment condition and by grade level. In spring 2014, teachers predicted their target students' responses for 504 students out of the 622 grade 1 and 2 students in the analytic sample who completed the MPAC interview. In spring 2015, teachers predicted their target students' responses for 785 students out of the 856 grades 1 and 2 students in the analytic sample who completed the MPAC interview. To our knowledge, any missing data are missing at random and not related to teacher or student characteristics or to the teacher judgment constructs of interest.

This sample is referred to in the present report as the *predicted analytic sample*. The MPAC student sample is representative of the entire project's larger student sample (Schoen, LaVenia, Champagne, & Farina, 2016; Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016). The predicted analytic sample of students included a diverse range of socioeconomic and ethnic characteristics. The sample demographics for the predicted analytic sample are similar to those of the larger student sample.

2.1.4. Random Selection of Students for Interviews

Interviews were conducted with a stratified random sample of up to four students from each participating teacher's classroom. The goal was to include two boys and two girls in the interview sample from each teacher's class. To maintain a balanced sample within each classroom with respect to student gender, gender was used as the first stratum. Student gender data were provided by the school districts. The second stratum involved splitting the class by achievement level on the basis of the fall student mathematics test (Schoen et al., 2016a, 2016b). A random number was assigned to each student, and the sample was sorted by gender, achievement level, and random number.

For each classroom, the median achievement level on the fall mathematics test was determined, and students who achieved at or below the classroom median constituted the lower fall test stratum and those who achieved above it constituted the upper fall test stratum. The class roster was then divided into four subcategories: upper fall test boys, lower fall test boys, upper fall test girls, and lower fall test girls. A primary and an alternate student were selected from each subcategory on the basis of the random number. The highest random number designated the primary student; the second highest the alternate. Alternate students were only called upon to be interviewed in instances where the primary student was absent or did not consent to be interviewed.

The MPAC interviews were conducted in spring 2014 and 2015 with students who completed the fall mathematics tests for the respective school year. In 2014, interviews were conducted only with students who had completed the fall 2013 test and had positive parental consent to be video recorded. In 2015, out of concerns that this latter criterion could introduce potential sampling bias, the students were drawn at random for the MPAC interview from the pool of students who had positive parental consent for participation in the study and completed the fall 2014 mathematics test. Students without parental consent to be video recorded remained in the MPAC interview sample for spring 2015.

Students without parental consent for a video-recorded interview (or who did not consent to be video recorded) in spring 2015 were interviewed while an observer recorded data on a data-collection sheet in real time as the interviewer conducted the interview and separately recorded data. The two sets of data were recorded independently and were used to permit examination of interrater reliability. After the separate coding was documented and the interview concluded, the interviewer and observer then resolved any discrepancies or missing data.

Table 3 offers a snapshot of the number of students interviewed for the teachers in the 2014 and 2015 sample, respectively. The goal to interview four target students from each participating teacher's classroom was largely achieved, especially in spring 2015. In a few instances, situations out of the researchers' control prevented all four of the randomly selected students from being interviewed. One teacher in the predicting analytic sample of 2014 and one in that of 2015 gave predictions for more than four students, because they were teacher of record for two different classes of students and gave predictions for interviewed students in both groups. These discrepancies occurred with similar frequencies for teachers in the treatment and control conditions.

Table 3. Number of Students Interviewed for Each Teacher by Year

Year	Number of teachers	Number of students per teacher
2014	95	4
	29	3
	11	2
	10	1
	1	5
2015	190	4
	3	3
	1	2
	2	1
	1	6

Interrater agreement checks for video-taped interviews were conducted on the basis of real-time observer data and observer data generated through a stratified random sample of videos. The proportions of interrater agreement on correctness for the overall interviews were .96 and .97 for the 2014 and 2015 MPAC interviews, respectively (Schoen, LaVenia, Champagne, & Farina, 2016; Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016). The interviewers were not made aware of the treatment condition of the school (or students), nor were they aware of the students' achievement on the fall mathematics test.

2.2. Measures

2.2.1. Fall Student Mathematics Tests: Grade 1 and Grade 2

Students with consent to participate in the study completed a written, group-administered mathematics test at the beginnings of the 2013–14 and 2014–15 school years. The tests were delivered to participating schools the week before students returned to school for the year. Teachers were asked to administer the tests within the first two weeks of the school year. Along with class rosters, tests were retrieved by members of the evaluation team approximately 4–6 weeks after the beginning of the school year.

The fall student mathematics tests were designed to measure student ability to answer correctly questions related to counting, solving word problems, and performing computation involving addition or subtraction. The tests were designed to be aligned with the learning expectations in the Common Core State Standards for Mathematics (NGACBP & CCSSO, 2010). The content and format of items and scales were reviewed by experts in mathematics and mathematics education. Data were modeled by means of a second-order factor-analysis model with Math as the higher-order factor and Counting, Word Problems, and Computation as three lower-order factors. The test forms at the two grade levels were not vertically scaled. The reliabilities of the fall mathematics test scales were determined from a composite reliability estimate for the second-order Math factor and ordinal forms of Cronbach's α for the subscales. The grade 1 Math composite reliability estimates were .84 and .88 for the fall 2013 and fall 2014 tests, respectively. The grade 2 Math composite reliability estimates were .89 and .91 for the fall 2013 and fall 2014 tests, respectively (Schoen et al., 2016a, 2016b).

2.2.2. Mathematics Performance and Cognition (MPAC) Interview

All of the items on the TAPSS interview were drawn from the MPAC interview (Schoen, LaVenia, Champagne, & Farina, 2016; Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016). Focused on the domains of number, operations, and equality, the Mathematics Performance and Cognition (MPAC) student interview was designed (a) to measure student achievement in mathematics and (b) to gather information about the strategies students use to solve the mathematics problems. The MPAC interview consists of a series of mathematics problems that the students are asked to solve in a one-on-one interview setting. The development process for the MPAC interviews involved expert review that verified the alignment of the content of the interview with current research and with fundamentally important ideas in first- and second-grade mathematics that are consistent with the content of the Common Core State Standards for Mathematics (NGACBP & CCSSO, 2010).

The MPAC interview uses a semistructured format. The interviewer poses a fixed set of problems to the student, observes how the student solves the problems, asks the students to report the strategies they used, and records the students' responses. The sequence and wording of the general instructions and the mathematics problems are designed to be presented in the same order and spoken exactly from the interviewer's script. Subsequent follow-up questions vary and depend upon the interviewer's ability to perceive and understand the student's strategy as well as the student's ability to demonstrate or articulate how he or she arrived at the given answer. The interview lasts approximately 45 minutes (range 30 to 60 minutes; Schoen, LaVenia, Champagne, & Farina, 2016; Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016).

The interviews were conducted by a team of research faculty with mathematics teaching experience and graduate students in mathematics education. Interviewer training occurred in several phases over a period of approximately 6 weeks. Each interview was video recorded in 2014. In 2015, the interviews were video recorded for students with video-recording consent, and answers were recorded on paper by an observer for students without parental consent to video record. The video recordings of a stratified random sample of 79 interviews in 2014 and of 210 interviews in 2015 were also coded by a separate trained reviewer as a check for consistency among interviewers of the implementation of the protocol and coding of data. The overall percentages of interrater agreement in determining whether students provided correct or incorrect answers were .96 and .99 in the 2014 and 2015 samples, respectively (Schoen, LaVenia, Champagne, & Farina, 2016; Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016).

2.2.3. The 2014 Teacher Ability to Predict Student Success (TAPSS) Measure

Items used in the TAPSS assessment were selected from MPAC interview items. Table 4 represents the list of four tasks used for the TAPSS measure during spring 2014. The same tasks were used to assess teacher judgment accuracy for both first and second grade teachers.

Table 4. Spring 2014 TAPSS Assessment Items

Item No.	Grade 1	Grade 2
Item 1	5 + ☐ = 13	5 + ☐ = 13
Item 2	6 + 3 = ☐ + 4	6 + 3 = ☐ + 4
Item 3	6 + 5 = ☐	6 + 5 = ☐
Item 4	4 + 8 = ☐	4 + 8 = ☐

Teachers were asked to predict student answers for each of the items detailed in Table 4 in 2014 and in Table 5 in 2015. In both years, the teachers were asked to predict whether each target student would solve each item correctly or incorrectly. Generally, teachers were asked to complete a prediction sheet for the target students the same day of the interview, shortly after students were interviewed. For the few times the teachers were absent or did not get a chance to fill out the prediction sheet when the students were interviewed, a researcher returned to the school within two weeks of the interview to give those teachers the opportunity to fill out the prediction sheet. In an effort to maintain data integrity, a member of the research team accompanied the teacher while he or she was in possession of the prediction sheet. The teacher prediction sheets are provided in Appendix B.

The spring 2014 teacher prediction sheet was identical for grade 1 and grade 2 teachers. For the three counting and computation items in spring 2014, teachers were also asked to predict whether the target student would know those facts at a recall level. This latter component was dropped from the 2015 TAPSS measure for reasons described below.

2.2.4. The 2015 Teacher Ability to Predict Student Success (TAPSS) Measure

Table 5 represents the tasks used for the TAPSS assessment during spring 2015. They differed from the set of items used in spring 2014 in several ways. The 2015 TAPSS instrument involved seven mathematics problems rather than four. By design, the problems selected for 2015 were more difficult for students (i.e., a lower percentage of students answered them correctly), and the items sampled a broader range of types of mathematics problems. The 2014 items primarily sampled computational ability with basic facts. The 2015 items sampled student knowledge and abilities related to basic addition facts, multidigit subtraction, the meaning of the equals sign, and solving word problems. The only mathematics problem present in both years of the TAPSS assessments was the equals sign item 6 + 3 = ☐ + 4 (Item 2 in spring 2014 and Item 3 in spring 2015).

As with the spring 2014 set of items, the same items were used in both grade levels in spring 2015, except for Item 7, which was unique to each grade level. The structure of the word problem used in item 7 was similar in the two grade levels, but the problem used in grade 1 involved smaller numbers and language that more explicitly described the succession of events in the story than the problem used in grade 2.

Table 5. Spring 2015 TAPSS Assessment Items

Item No.	Grade One	Grade Two
Item 1	10 = 7 + 3 [True or Not True]	10 = 7 + 3 [True or Not True]
Item 2	6 = 6 [True or Not True]	6 = 6 [True or Not True]
Item 3	6 + 3 = ☐ + 4	6 + 3 = ☐ + 4
Item 4	102 – 3 = ☐	102 – 3 = ☐
Item 5	21 – 19 = ☐	21 – 19 = ☐
Item 6	CDU(8, 15) James worked on his homework for 8 minutes. Courtney worked on her homework for 15 minutes. How many minutes longer did Courtney work on her homework than James?	CDU(8, 15) James worked on his homework for 8 minutes. Courtney worked on her homework for 15 minutes. How many minutes longer did Courtney work on her homework than James?
Item 7	JCU(15, 24) Caleb had 15 books on his shelf. Then he got some more books from the library and put them on his shelf. Now, he has 24 books on his shelf. How many books did Caleb get from the library?	JCU(25, 44) Aiden has collected 25 cards. He wants to collect 44 cards in total. How many more cards does Aiden need to collect?

Note. CDU = Compare difference unknown; JCU = Join change unknown (Carpenter et al., 2015).

The spring 2015 prediction sheets (Appendix B) were different for grade 1 and grade 2, to account for the differences in the last item on the assessment. For true or not true items, teachers were asked to predict whether the target student would answer with "True" or with "Not True."

In 2015, teachers were asked only to predict whether each target student would solve each item correctly or incorrectly for the other five tasks and were not asked any questions about student strategies. The question(s) about whether students would know the facts at a recall level were dropped for several reasons. First, we could not be sure that the teachers understood the intent of the question (which requires the teachers to have a common understanding of the phrase *recall level*). Second, it asks about teacher knowledge of their students' cognitive processes, which is conceptually different from the simpler question of whether a student would produce a correct answer. This component was dropped from the protocol to make space for more items and clarify the construct and the latent trait we were trying to measure with the TAPSS instrument (i.e., teacher judgment accuracy).

2.3. Data Analysis

Figures 1 and 2 present the structure of the spring 2014 and spring 2015 TAPSS data, respectively. Data from the two years were analyzed separately because of the differences in the predicting analytic sample and the items in the two assessments. The figures also illustrate the variables that will henceforth be associated with each level of the data.

We define the set $\mathbb{T}$ of size *n* to be the set of teachers who predicted one or more of their students' performances in a given year of the TAPSS assessment, where *k* is a teacher from this sample ($k \in \mathbb{T}$). We define $\mathbb{S}$ to be the set of students who took the MPAC assessment and for whom their teachers completed prediction sheets during one year of the TAPSS assessment. Most teachers had four students each in the predicted analytic sample, but some had other numbers. (See Table 3 for details.) The cardinality of the set was therefore $|\mathbb{S}| = \sum_{k=1}^{n} s_k$, where s_k is the number of students for whom the k^{th}

teacher predicted performance. We refer to one student from this sample as student *j* ($j \in \mathbb{S}$). Finally, we define $\mathbb{A}$ to be the set of all TAPSS item answers for all students and teachers for one year of the TAPSS instrument. Because some item-level data were missing for students in the predicted analytic sample (see Section 2.3.3 for more information about these cases), the cardinality of this set is $|\mathbb{A}| = \sum_{k=1}^{n}\sum_{j=1}^{s_k}(\alpha - missing_{jk})$, where α is the number of items used for TAPSS assessment for that year and $missing_{jk}$ is the number of TAPSS items with missing data for student *j*.

2.3.1. Data Entry Coding for Teacher Predictions and Student Success in Solving Problems

A teacher *k*'s prediction for student *j*'s performance on TAPSS item *i* was coded 0 if the teacher predicted that student *j* would solve the task incorrectly and was coded 1 if the teacher predicted that student *j* would solve the task correctly. Teacher prediction codes were recorded in the three-dimensional array T_{ijk}. Similarly, student *j*'s performance on TAPSS item *i* was coded 0 if the student solved the task incorrectly and was coded 1 if the student solved the task correctly. Students' performance codes were recorded in the three-dimensional array S_{ijk}.

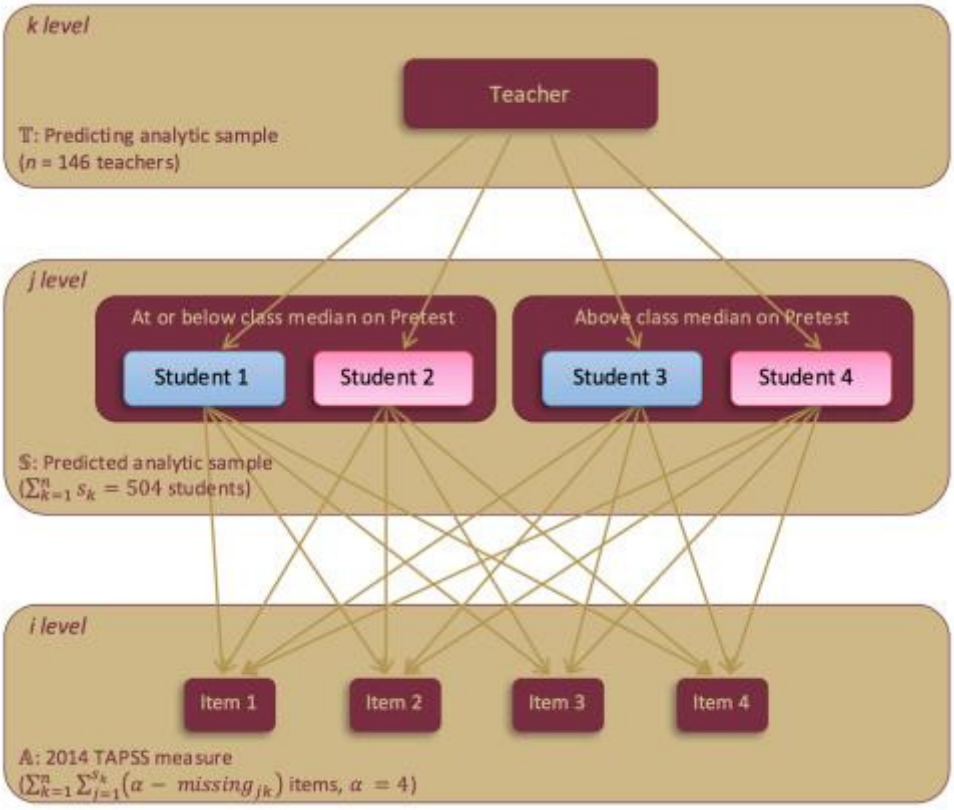

Figure 1. The structure of the 2014 TAPSS set of data with corresponding variable notations.

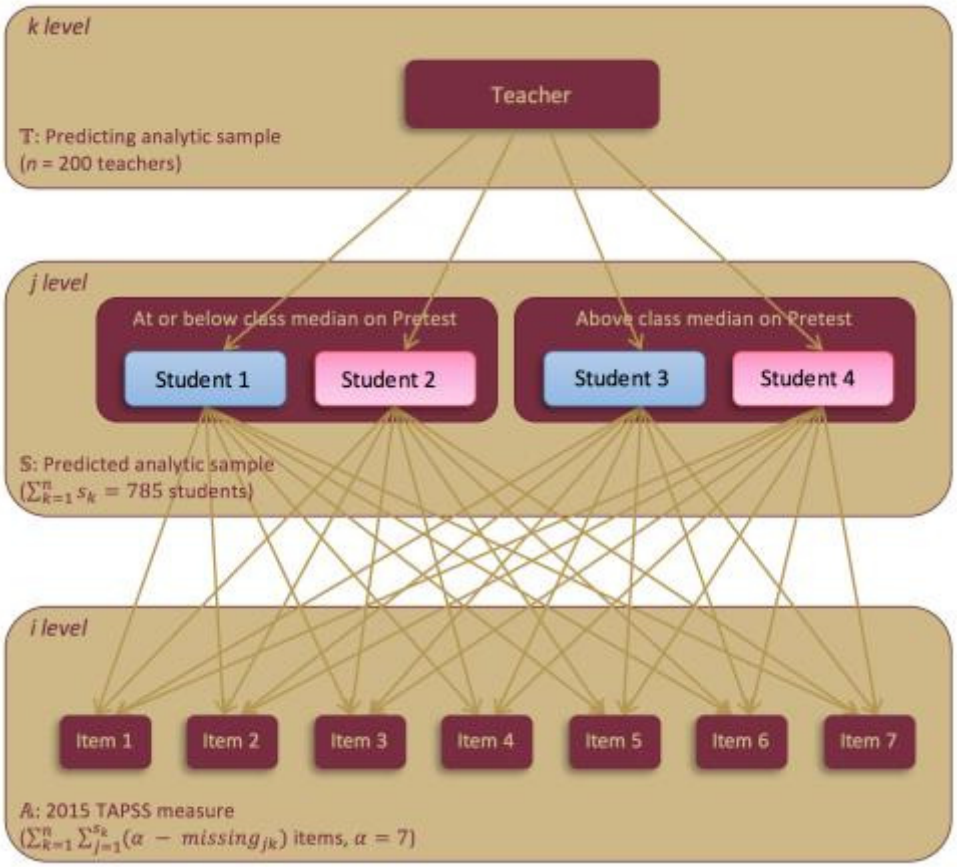

Figure 2. The structure of the 2015 TAPSS set of data with corresponding variable notations.

2.3.2. Coding Judgment Accuracy, Overprediction, and Underprediction

An individual teacher's ability to predict student success was calculated for each target student by comparison—for each TAPSS assessment item *i*—of teacher *k*'s prediction for student *j* (T_{ijk}) with student *j*'s actual performance (S_{ijk}). The binary nature of the predictions—accurate or inaccurate—and actual performance—correct or incorrect—created four possible outcomes, which were captured by means of four working variables. The variable named AC (i.e., accurate prediction with a correct answer) corresponds to the situation wherein the teacher predicted the student would solve the problem correctly and the student solved it correctly. The variable named AI (i.e., accurate prediction with an incorrect answer) corresponds to the situation where the teacher predicted the student would not solve the problem correctly and the student did not solve it correctly. Likewise, IC (i.e., inaccurate prediction with a correct answer) and II (i.e., inaccurate prediction with an incorrect answer) correspond to the remaining two possibilities. Table 6 displays these four situations and the way they were scored.

Table 6. Coding the Match between Teacher Prediction and Student Performance

Teacher Prediction T_{ijk}	Student Performance S_{ijk}	Accurate/ Correct AC_{ijk}	Accurate/ Incorrect AI_{ijk}	Inaccurate/ Correct IC_{ijk}	Inaccurate/ Incorrect II_{ijk}
1	1	1	0	0	0
0	0	0	1	0	0
0	1	0	0	1	0
1	0	0	0	0	1

For teacher *k*, student *j*, and item *i,* the comparison of teacher prediction with student performance was coded as 1 in the working variable that describes the comparison and as 0 in the three other variables as described in Table 6. If teacher *k* predicted that student *j* would solve item *i* correctly $(\text{i.e.}, T_{ijk} = 1)$, and student *j* solved item *i* correctly during the MPAC interview $(\text{i.e.}, S_{ijk} = 1)$, then $AC_{ijk} = 1$, $AI_{ijk} = 0,\ IC_{ijk} = 0,\ and\ II_{ijk} = 0$. If teacher *k* predicted student *j* would solve item *i* incorrectly $(\text{i.e.}, T_{ijk} = 0)$, and student *j* solved item *i* incorrectly $(\text{i.e.}, S_{ijk} = 0)$, this was coded as $AI_{ijk} = 1$. In the event of teacher *k* predicting student *j* would solve item *i* incorrectly $(\text{i.e.}, T_{ijk} = 0)$, and student *j* answering item *i* correctly $(\text{i.e.}, S_{ijk} = 1)$, this situation was coded as $IC_{ijk} = 1$. Finally, the case of teacher *k* predicting student *j* would solve interview item *i* correctly $(\text{i.e.}, T_{ijk} = 1)$, when student *j* answered item *i* incorrectly $(\text{i.e.}, S_{ijk} = 0)$, was coded as $II_{ijk} = 1$.

A teacher was considered to be *accurate* in his or her prediction for a task if the teacher's prediction matched the student's actual performance on that task during the MPAC interview, whether the student was predicted to answer correctly or to answer incorrectly. A teacher was considered to have *overpredicted* who predicted the target student would solve a task correctly and the student did not solve the task correctly (i.e., $II_{ijk} = 1$). A teacher was considered to have *underpredicted* who predicted the target student would solve a task incorrectly and the student solved the task correctly (i.e., $IC_{ijk} = 1$).

2.3.3. Procedures for Handling Missing Data

In spring 2014, teachers predicted their target students' responses for 504 students out of the 622 first and second grade students who participated in the MPAC interview. In spring 2015, the teachers predicted students' responses for 785 students out of 856. These figures reflect approximately 19% and 8% attrition from the intended samples in 2014 and 2015, respectively. Review of missing data did not identify patterns that would be expected to introduce bias. The missing data for the other 118 students from year one and 71 students from year two resulted primarily from absence of teachers from the building on the day the students were interviewed or from other reasons that teachers were unavailable to complete the form on that day or the days we revisited the school (e.g., had been called to attend unforeseen meetings).

During the interview, students were allowed to skip questions they did not wish to answer as per the MPAC interview protocol (Schoen, LaVenia, Champagne, & Farina, 2016; Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016). When students chose to skip a question, we assumed they had decided they could not solve it correctly.

Some students were also not presented with some of the questions on the basis of their previous responses, under the MPAC interview's *mercy rule*. The items in each subsection of the interview were generally sequenced from easiest to most difficult. The mercy rule allowed the interviewer to skip items

remaining in a given section when a student provided an incorrect answer (or no answer) on three consecutive items in the section (Schoen, LaVenia, Champagne, & Farina, 2016; Schoen, LaVenia, Champagne, Farina, & Tazaz, 2016). The underlying assumption of the *mercy rule* was that the student would not have solved the subsequent items in that section correctly. On the basis of this thinking, we considered both the skipped and mercy-ruled items as equivalent to the student's solving those items incorrectly and coded them accordingly.

Occasionally, students left some questions unanswered because the interviews ended prematurely because the school day ended or other activities interfered that were outside the control of the interviewer. In these cases, the items that had missing answers were considered missing (rather than incorrect) when we analyzed the percentages of accuracy and inaccuracy. We introduced the variable $missing_{jk}$ to account for the number of items with no data for student *j.* The denominator used in the calculation of teacher judgment accuracy, overprediction, or underprediction was reduced by this variable to lower the bias in teacher-level scores due to missing answers.[2] For example, during the 2015 interview, 12 students had one or more missing answers for these reasons, producing the small differences in the sample size across the various items. The sample size for each item is reported in the Results section.

2.3.4. Calculating Judgment Accuracy, Overprediction, and Underprediction by Item

After the data were entered and verified for accuracy, the first step in data analysis involved calculating the overall rate of teacher accuracy, overprediction, and underprediction for each item in the assessment. As shown in Equation 1, the overall rate of teacher judgment accuracy for each TAPSS item *i* was calculated by division of the total number of observed instances in which teachers correctly predicted the student would solve the item correctly by the total number of valid (i.e., nonmissing) data points for that item in the 2014 or 2015 sample, respectively.

$$Accuracy_i = \frac{\sum_{k=1}^{n}\sum_{j=1}^{s_k}(AC_{ijk} + AI_{ijk})}{\sum_{k=1}^{n}\sum_{j=1}^{s_k}(\alpha - missing_{jk})} \quad (1)$$

As shown in Equation 2, the overall rate of overprediction for each TAPSS item *i* was calculated by division of the total number of observed instances in which teachers incorrectly predicted the student would produce a correct answer by the total number of valid data points for that item.

$$Overprediction_i = \frac{\sum_{k=1}^{n}\sum_{j=1}^{s_k} II_{ijk}}{\sum_{k=1}^{n}\sum_{j=1}^{s_k}(\alpha - missing_{jk})} \quad (2)$$

As shown in Equation 3, the overall rate of underprediction for each TAPSS item *i* was calculated by division of the total number of observed instances in which teachers incorrectly predicted the student would not produce a correct answer by the total number of valid data points for that item.

$$Underprediction_i = \frac{\sum_{k=1}^{n}\sum_{j=1}^{s_k} IC_{ijk}}{\sum_{k=1}^{n}\sum_{j=1}^{s_k}(\alpha - missing_{jk})} \quad (3)$$

[2]This adjustment was used because we are currently using a classical approach to scoring teacher judgment accuracy. Scoring procedures based on item response would not have the same type of bias due to this type of missing answers and would not need this adjustment.

2.3.5. Teacher Judgment Accuracy, Overprediction, and Underprediction for Individual Students

We calculated teachers' accuracy, overprediction, and underprediction for each individual student in the analytic sample. Teacher accuracy was calculated for each student within teacher *k* by division of the number of times the teacher accurately predicted student *j* would solve the item correctly (i.e., AC + AI) by the number of items for which a student response was recorded.

Because the calculation of teacher judgment accuracy uses a classical test-theory approach, it is susceptible to bias caused by missing data. In an attempt to counteract the downward bias that would be introduced by using an unadjusted α term in the denominator, we reduced α by the number of missing items for that individual student-teacher combination. Although this procedure does not necessarily create an unbiased score—because differences in difficulty of the various items are not included in this model—it does address the known downward bias introduced by missing data.

Equation 4 shows how teacher accuracy was calculated for each student, where $\alpha = 4$ for the spring 2014 version of the TAPSS assessment and $\alpha = 7$ for the spring 2015 version. The $Accuracy_{jk}$ variable therefore corresponds to a percentage of the number of times teacher *k* was able to predict student *j*'s performance accurately and therefore has values between 0 and 1.

$$Accuracy_{jk} = \frac{\sum_{i=1}^{\alpha}(AC_{ijk}+AI_{ijk})}{\alpha - missing_j} \quad (4)$$

Equation 5 shows how teacher overprediction was calculated for each student in the same manner as teacher accuracy. The $Overprediction_{jk}$ variable corresponds to a percentage of the number of times teacher *k* incorrectly predicted student *j* would produce a correct answer and has values between 0 and 1.

$$Overprediction_{jk} = \frac{\sum_{i=1}^{\alpha} II_{ijk}}{\alpha - missing_j} \quad (5)$$

Teacher underprediction was similarly calculated for each student, as seen in Equation 6. The $Underprediction_{jk}$ variable corresponds to a percentage of the number of times teacher *k* incorrectly predicted student *j* would produce an incorrect answer and has values between 0 and 1.

$$Underprediction_{jk} = \frac{\sum_{i=1}^{\alpha} IC_{ijk}}{\alpha - missing_j} \quad (6)$$

2.3.6. Teachers' Judgment Accuracy, Overprediction, and Underprediction Across Students in Their Own Classes

After calculating teacher judgment accuracy, overprediction, and underprediction for each individual student, we determined each teacher's judgment accuracy, overprediction, and underprediction rate. A teacher's judgment accuracy rate was found by calculation of the arithmetic mean of the (usually four) $Accuracy_{jk}$ student scores for that teacher. The formula for the teacher judgment accuracy calculation is shown in Equation 7.

$$Accuracy_k = \frac{\sum_{j=1}^{s_k} Accuracy_{jk}}{s_k} = \frac{1}{s_k}\sum_{j=1}^{s_k}\left[\frac{\sum_{i=1}^{\alpha}(AC_{ijk}+AI_{ijk})}{\alpha - missing_j}\right] \quad (7)$$

Similarly, a teacher's overprediction rate was found by calculation of the arithmetic mean of the (usually four) $Overprediction_{jk}$ student scores for that teacher. The formula for the teacher overprediction calculation is shown in Equation 8.

$$Overprediction_k = \frac{\sum_{j=1}^{s_k} Overprediction_{jk}}{s_k} = \frac{1}{s_k}\sum_{j=1}^{s_k}\left[\frac{\sum_{i=1}^{\alpha}(II_{ijk})}{\alpha - missing_j}\right] \quad (8)$$

A teacher's underprediction score was found by the same process, calculation of the arithmetic mean of the (usually four) $Underprediction_{jk}$ student scores for that teacher. The formula for the teacher underprediction calculation is shown in Equation 9.

$$Underprediction_k = \frac{\sum_{j=1}^{s_k} Underprediction_{jk}}{s_k} = \frac{1}{s_k}\sum_{j=1}^{s_k}\left[\frac{\sum_{i=1}^{\alpha}(IC_{ijk})}{\alpha - missing_j}\right] \quad (9)$$

As in the calculation of teacher judgment accuracy for individual students, we reduced α by the number of missing items for that individual student-teacher combination in an attempt to negate a downward bias that would otherwise be introduced. Similarly, although the target number of students for each teacher was four, fewer or more students per teacher were occasionally included, so the s_k term is usually four in both the 2014 and 2015 samples but is occasionally higher or lower. (See Table 3 for more information.)

2.3.7. Determining Overall Judgment Accuracy, Overprediction, and Underprediction

To determine the overall rate of teacher judgment accuracy, overprediction, and underprediction in the predicting analytic sample, we calculated the sample mean for each of the three factors, $Accuracy_{jk}$, $Overprediction_{jk}$, and $Underprediction_{jk}$. Overall teacher judgment accuracy for the full sample was calculated by division of the sum of the teachers' student-specific accuracy scores by the total number of predictions made by the teachers in the predicting analytic sample as described in Equation 10. Overall teacher overprediction for the full sample was calculated by division of the sum of the teachers' student-specific overprediction scores by the total number of predictions made by the teachers as described in Equation 11. Overall teacher underprediction for the full sample was calculated by division of the sum of the teachers' student-specific underprediction scores by the total number of predictions made as described in Equation 12.

$$Overall\ Accuracy = \frac{\sum_{k=1}^{n}\sum_{j=1}^{s_k} Accuracy_{jk}}{\sum_{k=1}^{n} s_k} \quad (10)$$

$$Overall\ Overprediction = \frac{\sum_{k=1}^{n}\sum_{j=1}^{s_k} Overprediction_{jk}}{\sum_{k=1}^{n} s_k} \quad (11)$$

$$Overall\ Underprediction = \frac{\sum_{k=1}^{n}\sum_{j=1}^{s_k} Underprediction_{jk}}{\sum_{k=1}^{n} s_k} \quad (12)$$

For the purpose of further data exploration, overall teacher judgment accuracy, overprediction, and underprediction scores were calculated separately (1) within each grade level and (2) within each treatment condition.

3. Results

3.1. Item-level Contingency Tables

Data on task-specific teacher ability to predict student success for TAPSS item *i* is represented in contingency tables displaying the intersection of teacher prediction and student performance. The contingency tables are presented for each individual item, separately by grade level and year in which the items were administered to students.

3.1.1. Item-Level Contingency Tables for the 2014 Sample

Table 7a. Item 5 + ☐ = 13 in the 2014 Grade 1 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	66.4	15.0	81.4
Incorrect	5.5	13.1	18.6
Total	71.9	28.1	

Note. Sample represents 77 teachers and 274 of their grade 1 students. Overall percentage of accurate predictions was 79.5.

Table 7b. Item 5 + ☐ = 13 in the 2014 Grade 2 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	81.8	7.3	89.1
Incorrect	9.1	1.8	10.9
Total	90.9	9.1	

Note. Sample represents 69 teachers and 220 of their grade 2 students. Overall percentage of accurate predictions was 83.6.

Table 8a. Item 6 + 3 = ☐ + 4 in the 2014 Grade 1 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	10.6	35.0	45.6
Incorrect	4.0	50.4	54.4
Total	14.6	85.4	

Note. Sample represents 77 teachers and 274 of their grade 1 students. Overall percentage of accurate predictions was 61.0.

Table 8b. Item 6 + 3 = ☐ + 4 in the 2014 Grade 2 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	14.5	42.7	57.2
Incorrect	1.8	40.9	42.7
Total	16.3	83.6	

Note. Sample represents 69 teachers and 220 of their grade 2 students. Overall percentage of accurate predictions was 55.4.

Table 9a. Item 6 + 5 = ☐ in the 2014 Grade 1 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	93.1	5.8	98.9
Incorrect	0.7	0.4	1.1
Total	93.8	6.2	

Note. Sample represents 77 teachers and 277 of their grade 1 students. Overall percentage of accurate predictions was 93.5.

Table 9b. Item 6 + 5 = ☐ in the 2014 Grade 2 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	97.8	1.8	99.6
Incorrect	0.0	0.4	0.4
Total	97.8	2.2	

Note. Sample represents 69 teachers and 227 of their grade 2 students. Overall percentage of accurate predictions was 98.2.

Table 10a. Item 4 + 8 = ☐ in the 2014 Grade 1 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	90.9	8.3	99.2
Incorrect	0.4	0.4	0.8
Total	91.3	8.7	

Note. Sample represents 77 teachers and 276 of their grade 1 students. Overall percentage of accurate predictions was 91.3.

Table 10b. Item 4 + 8 = ☐ in the 2014 Grade 2 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	95.2	3.1	98.3
Incorrect	1.3	0.4	1.7
Total	96.5	3.5	

Note. Sample represents 69 teachers and 227 of their grade 2 students. Overall percentage of accurate predictions was 95.6.

3.1.2. Item-Level Contingency Tables for the 2015 Sample

Table 11a. Item 10 = 7 + 3 [True or Not True] in the 2015 Grade 1 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	48.6	36.1	84.7
Incorrect	4.3	11.0	15.3
Total	52.9	47.1	

Note. Sample represents 106 teachers and 418 of their grade 1 students. Overall percentage of accurate predictions was 59.6.

Table 11b. Item 10 = 7 + 3 [True or Not True] in the 2015 Grade 2 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	61.4	31.4	84.7
Incorrect	3.3	3.9	7.2
Total	64.7	35.3	

Note. Sample represents 94 teachers and 363 of their grade 2 students. Overall percentage of accurate predictions was 65.3.

Table 12a. Item 6 = 6 [True or Not True] in the 2015 Grade 1 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	62.8	27.0	89.8
Incorrect	5.3	5.0	10.3
Total	68.0	32.0	

Note. Sample represents 106 teachers and 419 of their grade 1 students. Overall percentage of accurate predictions was 67.8.

Table 12b. Item 6 = 6 [True or Not True] in the 2015 Grade 2 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	77.4	17.4	94.8
Incorrect	3.6	1.7	5.3
Total	81.0	19.1	

Note. Sample represents 94 teachers and 363 of their grade 2 students. Overall percentage of accurate predictions was 79.1.

Table 13a. Item 6 + 3 = ☐ + 4 in the 2015 Grade 1 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	12.9	41.4	54.3
Incorrect	2.4	43.3	45.7
Total	15.2	84.7	

Note. Sample represents 106 teachers and 420 of their grade 2 students. Overall percentage of accurate predictions was 56.2.

Table 13b. Item 6 + 3 = ☐ + 4 in the 2015 Grade 2 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	16.4	40.6	57.0
Incorrect	2.2	40.9	43.1
Total	18.6	81.5	

Note. Sample represents 94 teachers and 365 of their grade 2 students. Overall percentage of accurate predictions was 57.3

Table 14a. Item 102 – 3 = ☐ in the 2015 Grade 1 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	37.3	18.8	56.1
Incorrect	14.7	29.3	44.0
Total	52.0	48.1	

Note. Sample represents 106 teachers and 416 of their grade 1 students. Overall percentage of accurate predictions was 66.6.

Table 14b. Item 102 – 3 = ☐ in the 2015 Grade 2 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	56.6	16.9	73.5
Incorrect	14.1	12.4	26.5
Total	70.7	29.3	

Note. Sample represents 94 teachers and 362 of their grade 2 students. Overall percentage of accurate predictions was 69.0.

Table 15a. Item 21 – 19 = ☐ in the 2015 Grade 1 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	35.6	38.9	74.5
Incorrect	7.9	17.6	25.5
Total	43.5	56.5	

Note. Sample represents 106 teachers and 416 of their grade 1 students. Overall percentage of accurate predictions was 53.2.

Table 15b. Item 21 – 19 = ☐ in the 2015 Grade 2 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	60.5	30.1	90.6
Incorrect	4.1	5.3	9.4
Total	64.6	35.4	

Note. Sample represents 94 teachers and 362 of their grade 2 students. Overall percentage of accurate predictions was 65.8.

Table 16a. Item CDU(8, 15) in the 2015 Grade 1 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	31.4	23.1	54.5
Incorrect	10.5	35.0	45.5
Total	41.9	58.1	

Note. CDU = Compare difference unknown (Carpenter et al., 2015). Sample represents 106 teachers and 420 of their grade 1 students. Overall percentage of accurate predictions was 66.4.

Table 16b. Item CDU(8, 15) in the 2015 Grade 2 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	52.6	14.3	66.9
Incorrect	16.4	16.7	33.1
Total	69.0	31.0	

Note. CDU = Compare difference unknown (Carpenter et al., 2015). Sample represents 94 teachers and 365 of their grade 2 students. Overall percentage of accurate predictions was 69.3.

Table 17a. Item JCU(15, 24) in the 2015 Grade 1 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	29.4	35.4	64.8
Incorrect	6.0	29.2	35.2
Total	35.4	64.6	

Note. JCU = Join change unknown (Carpenter et al., 2015). Sample represents 106 teachers and 415 of their grade 1 students. Overall percentage of accurate predictions was 58.6.

Table 17b. Item JCU(15, 24) in the 2015 Grade 2 Sample

	Student performance		
Teacher prediction	Correct	Incorrect	Total
Correct	32.7	30.8	63.5
Incorrect	8.2	28.3	36.5
Total	40.9	59.1	

Note. JCU = Join change unknown (Carpenter et al., 2015). Sample represents 94 teachers and 364 of their grade 2 students. Overall percentage of accurate predictions was 61.0.

3.1.3. Summary of Item-level Results

Table 18. Summary of Observed Proportions of Student Correctness, Teacher Predictions of Correctness, Accurate Predictions, Overpredictions, and Underpredictions for the Spring 2014 TAPSS Assessment Items, by Grade Level

	Grade 1					Grade 2				
Item	S	T	A	O	U	S	T	A	O	U
5 + ☐ = 13	71.9	81.4	79.5	15.0	5.5	90.9	89.1	83.6	7.3	9.1
6 + 3 = ☐ + 4	14.6	45.6	61.0	35.0	4.0	16.3	57.2	55.4	42.7	1.8
6 + 5 = ☐	93.8	98.9	93.5	5.8	0.7	97.8	99.6	98.2	1.8	0.0
4 + 8 = ☐	91.3	99.2	91.3	8.3	0.4	96.5	98.3	95.6	3.1	1.3

Note. S = Percentage of students solving the problem correctly; T = Percentage of teacher predictions that individual students would solve the problem correctly; A = Percentage of accurate predictions; O = Percentage of overpredictions; U = Percentage of underpredictions.

Overall, teachers predicted higher levels of student performance than the interviewers observed. In almost every case, the percentage of teachers who predicted students would solve the problems correctly was higher than the observed percentage of students who did so.

At both grade levels, more than 90% of students solved two of the four items in the TAPSS assessment correctly, and more than 98% of teachers predicted their students would solve those two problems correctly. The $Accuracy_i$ rate for those two items was over 90% at both grade levels. The result was very little variation in the predictions data.

The lowest rate of student success in correctly solving the problem (and the most variation in the data) was found in the nontraditional format (i.e., 6 + 3 = ☐ + 4). Overall, teachers seemed to be aware that this item would be more difficult than the other items, because less than half of the teachers predicted their students would solve this item correctly. Even though the prediction rate was half of that of the other items, the gap between the student success rate and the rate of teacher predictions of success was greatest on this item, suggesting that teachers also have less accurate knowledge (and higher overprediction) of their students' success on this item than on the other three items.

Table 19. Summary of Observed Proportions of Student Correctness, Teacher Predictions of Correctness, Accurate Predictions, Overpredictions, and Underpredictions for the Spring 2015 TAPSS Assessment Items, by Grade Level

	Grade 1					Grade 2				
Item	S	T	A	O	U	S	T	A	O	U
10 = 7 + 3	52.9	84.7	59.6	36.1	4.3	64.7	92.8	65.3	31.4	3.3
6 = 6	68.0	89.8	67.8	27.0	5.3	81.0	94.8	79.1	17.4	3.6
6 + 3 = ☐ + 4	15.2	54.3	56.2	41.4	2.4	18.6	57.0	57.3	40.6	2.2
102 – 3 = ☐	52.0	56.1	66.6	18.8	14.7	70.7	73.5	69.0	16.9	14.1
21 – 19 = ☐	43.5	74.5	53.2	38.9	7.9	64.6	90.9	65.8	30.1	4.1
CDU (8, 15)	41.9	54.5	66.4	23.1	10.5	69.0	66.9	69.3	14.3	16.4
JCU (15, 24)	35.4	64.8	58.6	35.4	6.0	–	–	–	–	–
JCU (25, 44)	–	–	–	–	–	40.9	63.5	61.0	30.8	8.2

Note. Items 1 and 2 involved having students say whether the equations were true or not true. CDU = Compare difference unknown; JCU = Join change unknown (Carpenter et al., 2015). S = Percent of students solving the problem correctly; T = Percent of teacher predictions that individual students would solve the problem correctly; A = Percent of accurate predictions; O = Percent of overpredictions; U = Percent of underpredictions.

Overall, variation in student performance and teacher predictions in the was much higher 2015 data than in the 2014 data. The phenomenon observed in the 2014 data set involving the equals-sign item (6 + 3 = ☐ + 4) was repeated in all three equals-sign items in the 2015 data; teachers overwhelmingly overpredicted their students' achievement on these items. Although student performance on the 2015 items was much lower than their performance on basic facts-type items in the 2014 TAPSS instrument, the rate of teacher predictions of success was only slightly lower on two of the equals sign items than on the items with student success rates above 90 percent in 2014. This result suggests that items involving asking teachers about their students' understanding of the equals sign may discriminate among teachers' levels of knowledge of their students' mathematical thinking or abilities.

3.2. Teacher Prediction of Individual Students' Performance

The graphs in this section illustrate the distribution of the teachers' judgment accuracy, overprediction, and underprediction rates for their individual students. Teachers' prediction scores were constrained by the number of items on the TAPSS assessment that students solved, which was generally four items for 2014 TAPSS and seven items for 2015 TAPSS. To account for all possible cases, the graphs for the 2014 sample have five bins and the graphs for the 2015 sample have eight bins. The prediction accuracy for individual students is higher in the 2014 sample than in the 2015 sample.

3.2.1. Teacher Prediction of Individual Students' Performance in the Spring 2014 Sample

Figures 3, 4, and 5 display the distribution of the $Accuracy_{jk}$, $Overprediction_{jk}$, and $Underprediction_{jk}$ scores, respectively, for the 2014 sample of teachers and students. Teachers were asked to predict their students' performance for four items.

Figure 3 provides a visualization of the distribution of the $Accuracy_{jk}$ scores in the 2014 sample as calculated by Equation 4. Approximately half of the students' performances were accurately predicted by their teachers for three of the four items in the measure. Almost as many students' performances were accurately predicted by their teachers for all four items. These results show that teachers

demonstrated high accuracy in predicting student performance for the four items used in the 2014 TAPSS assessment.

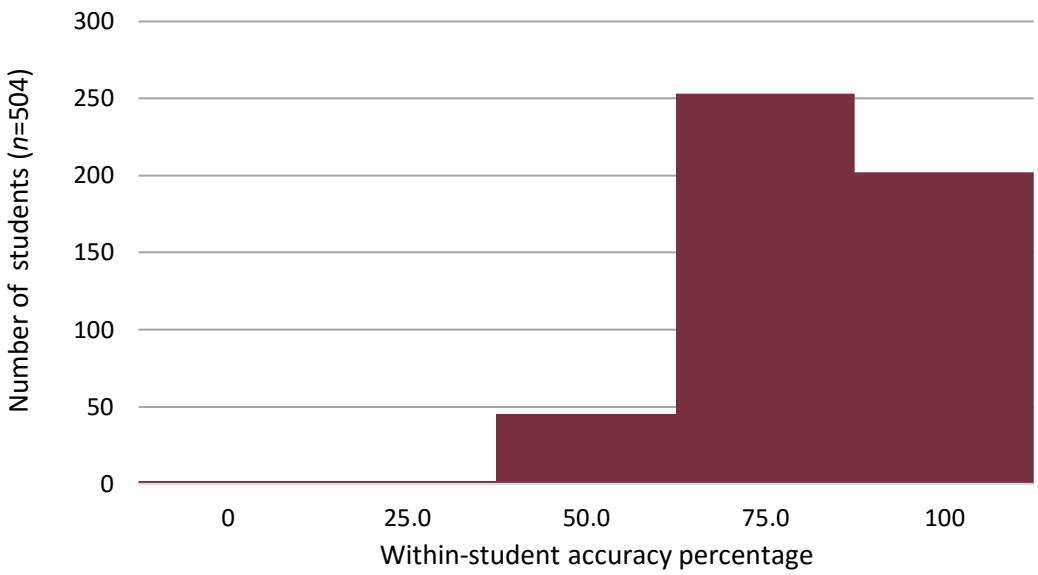

Figure 3. Distribution of 2014 teacher sample accuracy for individual students (i.e., $Accuracy_{jk}$).

Figure 4 offers a visualization of the distribution of the $Overprediction_{jk}$ scores in the 2014 sample as calculated by Equation 5. Approximately half of the sample students' performances were not overpredicted by their teachers for any of the four items. Almost as many students' performances were overpredicted by their teachers on one of the four items. On the basis of the item data seen in Tables 7a–10b, as well as in Table 18, most of these overpredictions occurred on TAPSS item 6 + 3 = ☐ + 4.

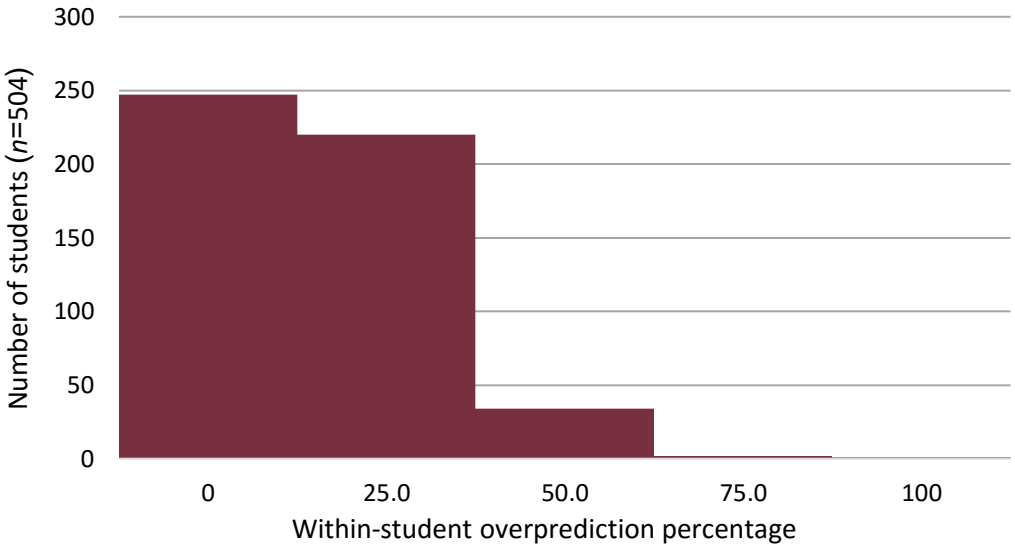

Figure 4. Distribution of 2014 teacher sample overprediction for individual students (i.e., $Overprediction_{jk}$).

Figure 5 provides a visualization of the distribution of the *Underprediction*$_{jk}$ scores in the 2014 sample as calculated by Equation 6. Underprediction was extremely rare; the overwhelming majority of student cases had zero instances of underprediction. Only about 10% of students' performances were underpredicted for one of the four items. These results show that teachers did not tend to underpredict student performance for the four items used in the 2014 TAPSS assessment.

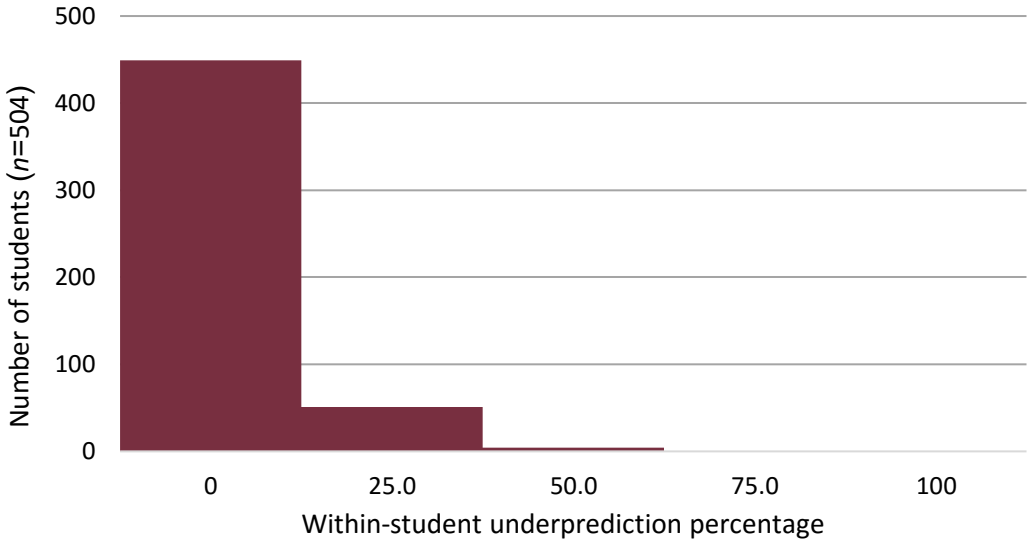

Figure 5. Distribution of 2014 teacher sample underprediction for individual students (i.e., Underprediction$_{jk}$*).*

3.2.2. Teacher Prediction of Individual Students' Performances in the Spring 2015 Sample

Figures 6, 7, and 8 display the distribution of the *Accuracy*$_{jk}$, *Overprediction*$_{jk}$, and *Underprediction*$_{jk}$ scores, respectively, for the 2015 sample. Teachers were asked to predict their students' performance for seven items in the spring 2015 version of the TAPSS instrument. For the few students who had missing data for one or more items, the rate of teacher prediction scores was calculated on the basis of the items for which data were collected (or available). In these cases, the teacher prediction rates for that student may have not have been a multiple of $\frac{1}{7}$ and were counted in the bins with closest percentage value. For example, teacher judgment accuracy for a teacher accurately predicting five out of six items their student solved (i.e. Accuracy$_{jk}$ = .833) was counted in the 85.7% bin, which corresponds to students whose performance was accurately predicted for six out of seven items. As described in Section 2.3.3, a few instances of missing data of this variety arose; only 12 instances of missing item-level data arose in the predicted analytic sample in 2015.

Figure 6 provides a visualization of the distribution of the *Accuracy*$_{jk}$ scores in the 2015 sample as calculated by Equation 4. The distribution has a unimodal, bell shape. Approximately half of the students' performances were accurately predicted for four or five of the seven items. Almost 10% of the students' performances were accurately predicted by their teachers for all seven items of the 2015 TAPSS assessment. These results show that teachers demonstrated a large range in accurately predicting student performance for the seven items used in the 2015 TAPSS assessment. They were more likely to predict at least four items accurately than three or fewer.

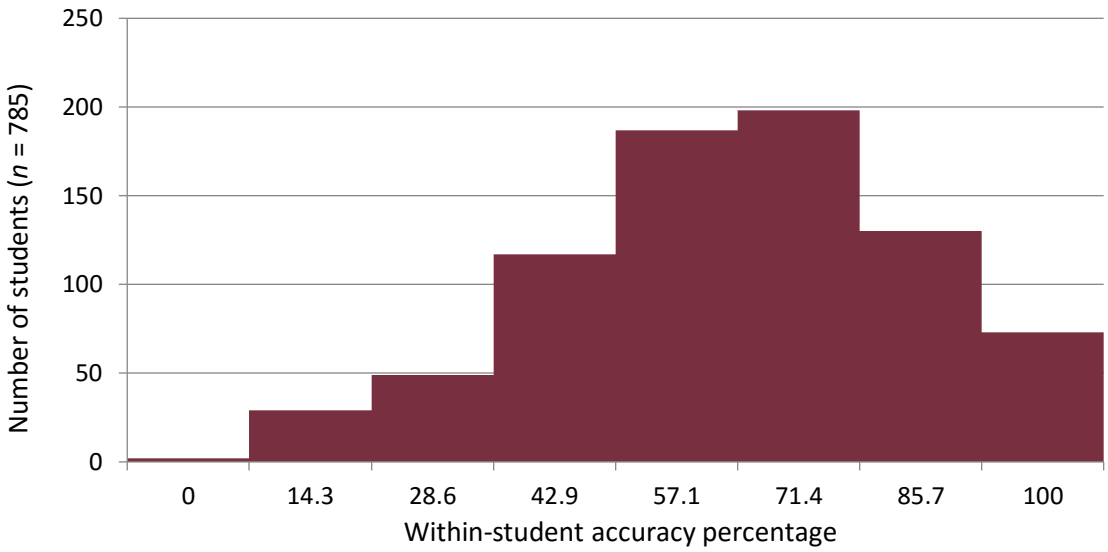

Figure 6. Distribution of 2015 teacher sample accuracy for individual students (i.e., $Accuracy_{jk}$).

Figure 7 offers a visualization of the distribution of the $Overprediction_{jk}$ scores in the 2015 sample as calculated by Equation 5. The distribution has a unimodal, bell shape and is skewed right. There were no instances of overprediction for fifteen percent of the students in the sample. One, two, or three instances of overprediction occurred for 70% of the students in the sample. Fifteen percent of the students in the sample were overpredicted four or more times.

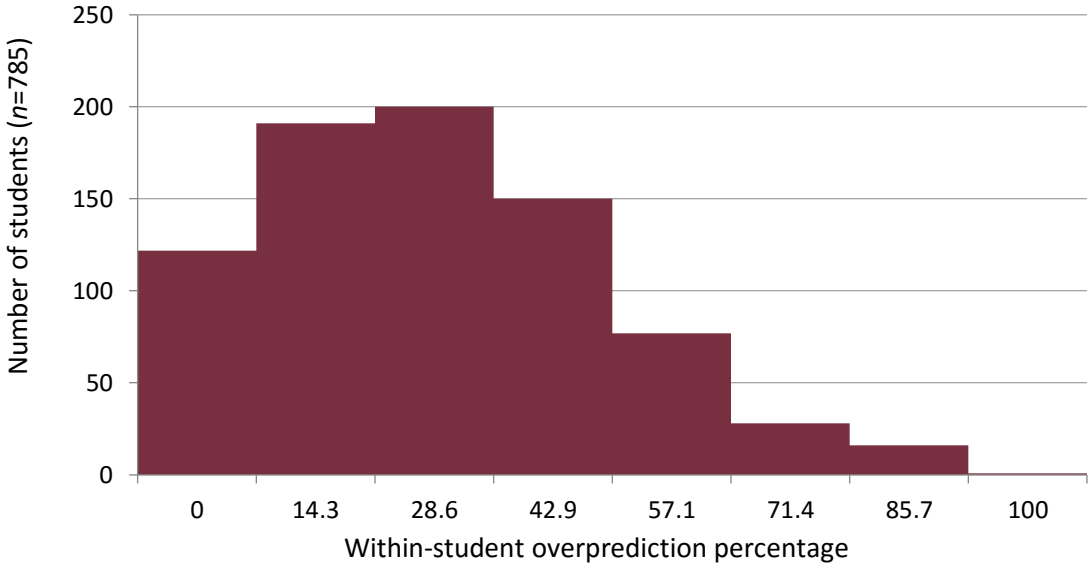

Figure 7. Distribution of 2015 teacher sample overprediction for individual students (i.e., $Overprediction_{jk}$).

Figure 8 provides a visualization of the distribution of the *Underprediction*$_{jk}$ scores in the 2015 sample as calculated by Equation 6. The distribution is severely skewed right. Underprediction was not observed within individual students for almost two-thirds of the student sample. Underprediction was observed for exactly one of the seven items on approximately one-fourth of the student sample. These results show that teachers did not tend to underpredict their students' performance on the seven items used in the 2015 TAPSS assessment.

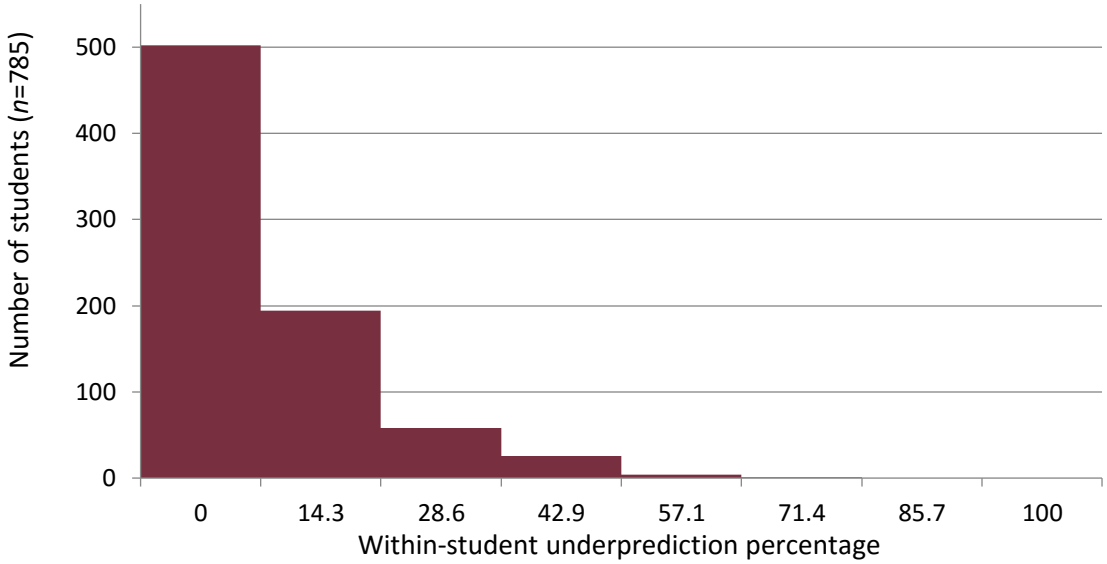

Figure 8. Distribution of 2015 teacher sample underprediction for individual students (i.e., Underprediction$_{jk}$*).*

3.3. Distribution of Individual Teachers' Mean Predictions

The graphs in this section illustrate the distribution of how accurate the teachers were in their predictions. Each teacher's judgment accuracy, overprediction, and underprediction rates were calculated as the arithmetic means of that teachers' predictions for his or her target students. These graphs are different from those in the previous section in that they show the overall predictions of a teacher to reveal whether that teachers' predictive accuracies, and therefore knowledge of their students' mathematic abilities, show a great range. The mean prediction accuracy for individual teachers is higher in the 2014 sample than in 2015.

3.3.1. Individual Teachers' Prediction of Their Students' Performance in the Spring 2014 Sample

Figures 9, 10, and 11 display the distribution of the *Accuracy*$_k$, *Overprediction*$_k$, and *Underprediction*$_k$ scores, respectively, for teachers in the 2014 sample.

Figure 9 offers a visualization of the distribution of the *Accuracy*$_k$ scores in the 2014 sample as calculated by Equation 7. Every teacher had a mean accuracy score greater than 50%. Most had a mean accuracy score greater than 75%. Teachers clearly demonstrated high accuracy in predicting student performance for the four items used in the 2014 TAPSS assessment.

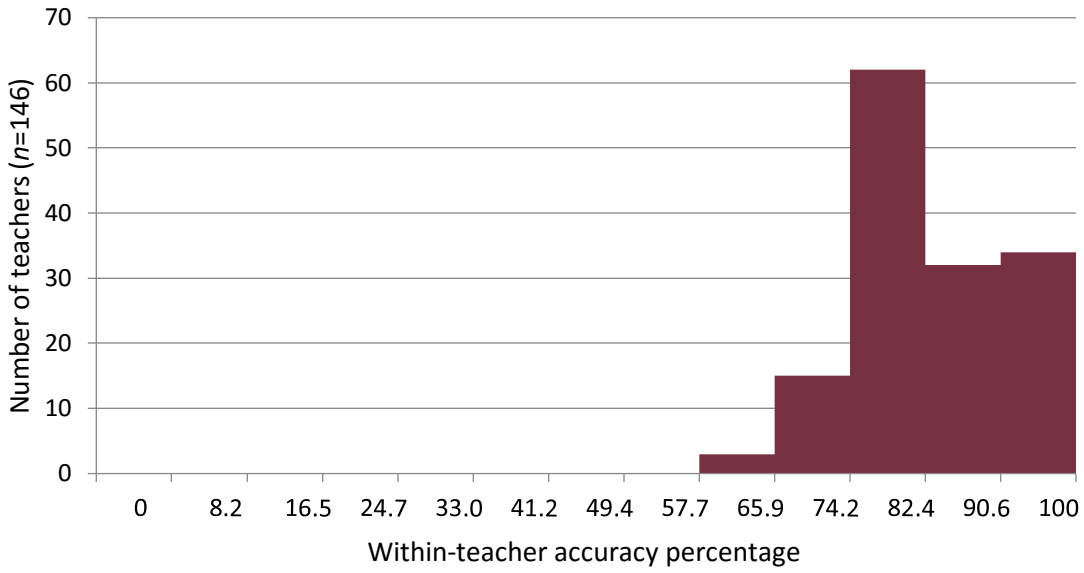

Figure 9. Distribution of 2014 sample mean accuracy for individual teachers (i.e., $Accuracy_k$).

Figure 10 offers a visualization of the distribution of the $Overprediction_k$ scores in the 2014 sample as calculated by Equation 8. The distribution is skewed left, and almost all teachers had a mean overprediction score of less than 30%. Almost 1% of the teachers in the 2014 sample did not overpredict their students' performance even once.

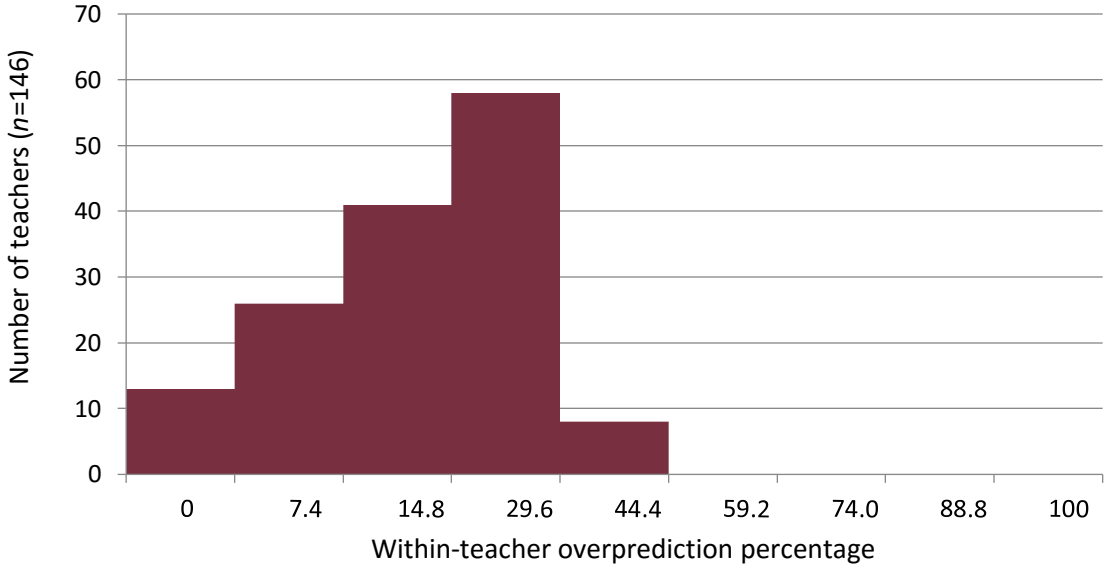

Figure 10. Distribution of 2014 sample mean overprediction for individual teachers (i.e., $Overprediction_k$).

Figure 11 offers a visualization of the distribution of the $Underprediction_k$ scores in the 2014 sample as calculated by Equation 9. The distribution is unimodal and heavily skewed right, where approximately two-thirds of the teachers did not underpredict their student performance on any of the four items. All of the teachers had underprediction scores under 20 percent based on the 2014 TAPSS assessment.

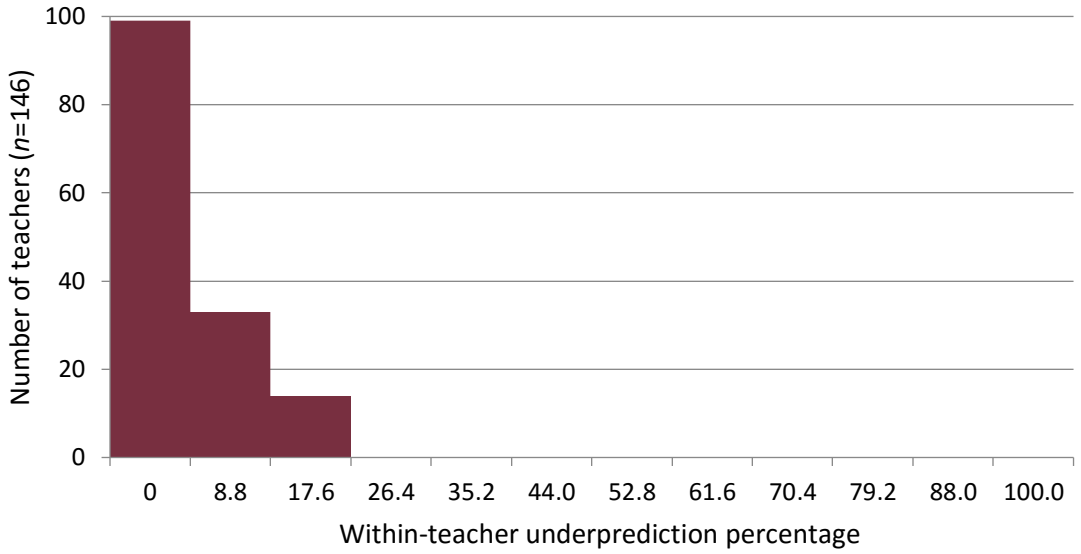

Figure 11. Distribution of 2014 sample mean underprediction for individual teachers (i.e., Underprediction$_k$).

3.3.2. Teachers' Predictions of Their Classes' Performance in the Spring 2015 Sample

Figures 12, 13, and 14 display the distribution of the Accuracy$_k$, Overprediction$_k$, and Underprediction$_k$ scores, respectively, for the 2015 sample.

Figure 12 offers a visualization of the distribution of the *Accuracy$_k$* scores in the 2015 sample as calculated by Equation 7. The distribution appears bell-shaped, and no teachers averaged 100% or less than 30% in accuracy. Most teachers had mean accuracy scores greater than 50%. Almost two-thirds of the teachers in the sample had accuracy scores between 53 and 75%. Although the accuracy rates are lower than in the 2014 sample (based on the four items in the 2014 TAPSS assessment), these results show that teachers demonstrated high rates of accuracy in predicting student performance for the seven items used in the 2015 TAPSS assessment.

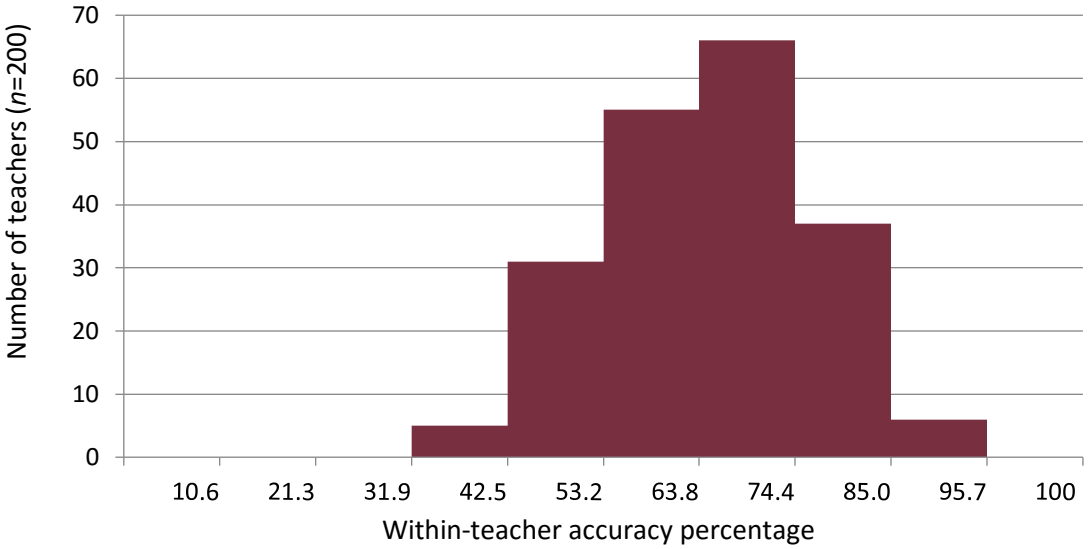

Figure 12. Distribution of 2015 sample mean accuracy for individual teachers (i.e., $Accuracy_k$).

Figure 13 provides a visualization of the distribution of the teacher-specific $Overprediction_k$ scores in the 2015 sample as calculated by Equation 8. The distribution appears bell-shaped. No teachers had average rates of overprediction rates greater than 70%, and some teachers did have very low rates of overprediction. Most teachers had overprediction scores below 50%. Approximately two-thirds of teachers had overprediction scores between 51 and 36%. In the aggregate, these results show that most teachers had a tendency to overpredict at least some of their students' performance for the seven items used in the 2015 TAPSS assessment.

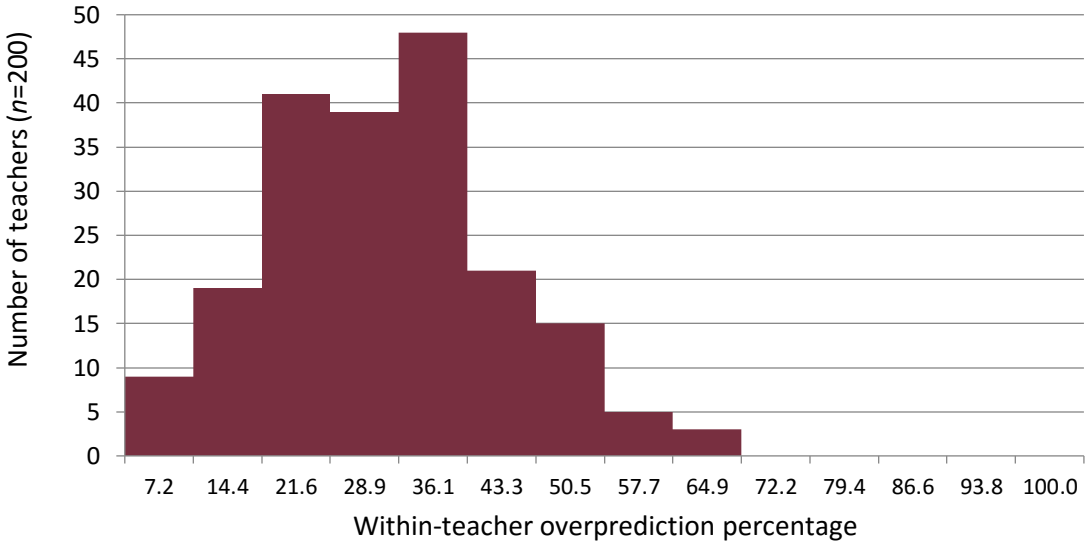

Figure 13. Distribution of 2015 sample mean overprediction for individual teachers (i.e., $Overprediction_k$).

Figure 14 provides a visualization of the distribution of the teacher-specific $Underprediction_k$ scores in the 2015 sample as calculated by Equation 9. The distribution is skewed right, and no teachers had mean rates of underprediction greater than 50%. Almost all of the teachers had average underprediction scores below 18%. Approximately two-thirds of teachers had mean underprediction

scores below 9%. In the aggregate, these results show that teachers were not prone to underpredict their students' performance for the seven items used in the 2015 TAPSS assessment.

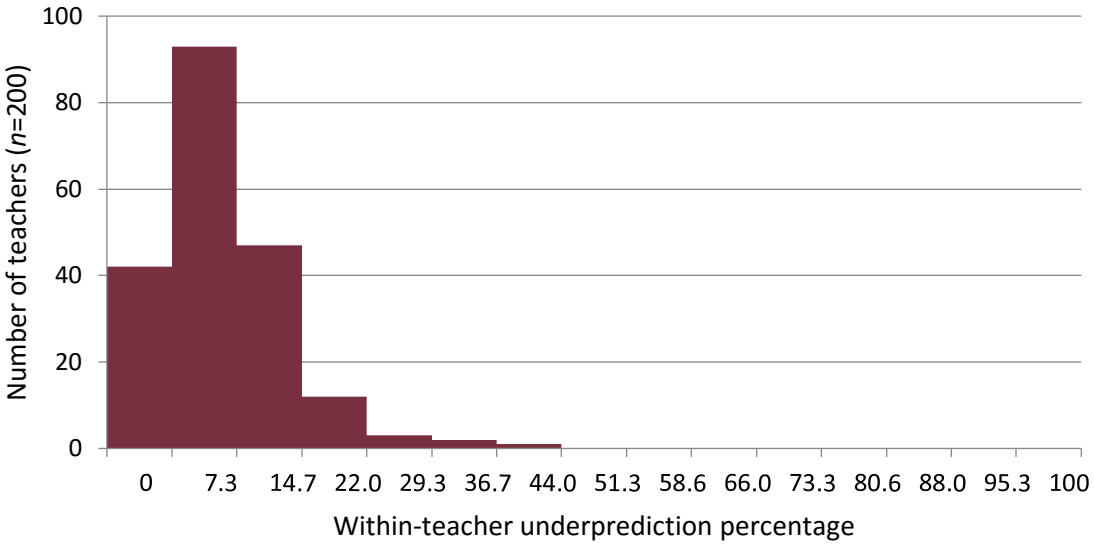

Figure 14. Distribution of 2015 sample mean underprediction for individual teachers (i.e., $Underprediction_k$).

3.4. Overall Percentages of Accuracy, Overprediction, and Underprediction

Tables 20 and 21 offer a big-picture snapshot of the overall accuracy, overprediction, and underprediction in the two years' samples. The overall prediction accuracy is higher in the 2014 sample than in 2015. The tables also break down these figures into subgroups by grade level and by treatment condition. We note that these descriptive statistics do not comprise a rigorous comparison of treatment and control. A careful analysis of treatment effect will need to account for the clustering of students within classroom in order to draw inferences with respect to the effect of grade level or treatment condition on teachers' ability to predict student performance.

Table 20. Spring 2014 Overall Percentage Accuracy, Overprediction, and Underprediction

Subgroup	Prediction accuracy		Overprediction		Underprediction	
	M	SD	M	SD	M	SD
Grade 1	81.4	18.0	15.9	17.4	2.7	8.3
Grade 2	83.3	15.5	13.4	14.7	3.2	9.0
Treatment	83.3	17.8	13.9	16.8	2.9	9.0
Control	81.4	16.2	15.5	15.9	3.0	8.4
Total sample	82.3	16.9	14.8	16.3	2.9	8.6

Table 21. Spring 2015 Overall Percentage Accuracy, Overprediction, and Underprediction

Subgroup	Prediction accuracy		Overprediction		Underprediction	
	M	SD	M	SD	M	SD
Grade 1	61.2	22.6	31.5	22.0	7.3	11.6
Grade 2	66.7	20.3	25.8	19.1	7.5	11.8
Treatment	66.0	21.7	26.2	19.6	7.8	12.4
Control	62.1	21.6	30.8	21.5	7.1	11.1
Total sample	63.7	21.7	28.9	20.8	7.4	11.7

4. Discussion

The TAPSS measure of teacher judgment accuracy is much more specific than other extant measures of this construct. The teachers predicted the success of individual students rather than of their whole classes. They predicted those students' success on individual items rather than their overall test score. The teachers were shown the test items before making their predictions. The TAPSS instrument contains fewer items and presents a smaller burden on the test takers than previously used instruments measuring a similar construct (cf. Carpenter et al., 1989; Gabriele et al., 2016). In the framework and language provided by Hoge and Coladarci (1989) and Südkamp et al. (2012), these design factors represent high levels of informed judgment, congruence, and rate of specificity.

Unlike Gabriele et al., we focused this study's data collection on observed success in producing correct answers and did not include other components such as confidence. The 2014 measure included a component related to teacher predictions of cognitive strategies, but concerns about whether teachers interpreted the notion of direct recall of facts consistently—especially the teachers in the control group schools—caused us to remove that component from the 2015 instrument. Curiously, the calibration of the TAPSS measurement from year one to year two resulted in overall percentage accuracy that fits in the range of the judgment accuracy found in the literature (cf. Hoge & Coladarci, 1989; Südkamp et al., 2012), although we were not aware of the work of Südkamp et al. at the time.

One important lesson of our work is the clear demonstration that the items affect the score. This effect is most clear in the 2014 sample. Teachers were highly accurate on the items involving basic number facts, but they were much less accurate on the item related to students' understanding of the meaning of the equals sign. This discrepancy resulted in very little variation in the accuracy, overprediction, and underprediction data for the 2014 sample. The 2015 sample showed much more variation, perhaps, in part, simply because it included more items (n = 7) than did the 2014 sample (n = 4). It is also due to a careful selection of items that (a) were more difficult for students, (b) covered a range of topics in the mathematics curriculum, and (c) represented topics about which teachers are often surprised to learn what their students do or do not know (e.g., meaning of the equals sign).

We are certain that the work to develop the items to measure teacher judgment accuracy benefited from the opportunity to administer it twice. The task-level analysis shows us how important choosing the proper items is for assessing judgment accuracy. After the first field experiment in spring 2014, a quick analysis of the data found very little variation in student responses, largely because of low difficulty (for the student sample) of most of the four items used. The students successfully produced correct answers, and the teachers knew they would. On the basis of these revelations, we revised the set of items in 2015 to include items that were more difficult (i.e., were solved correctly by a lower percentage of students) and more items overall. We also focused on teacher prediction of student success in solving the problem and discontinued measurement of teacher prediction of cognitive processes between the 2014 and 2015 iterations. Another key difference included the breadth of types of items in the spring 2015 predicted set of items. The spring 2015 item set included word problems, problems involving basic addition and subtraction facts, problems involving multidigit computation, and problems exploring student understanding of the meaning of the equals sign.

The level of difficulty of the set of items on the 2015 tests appears to be reasonably well calibrated for both the grade 1 and grade 2 samples. The overall rate of accurate predictions was higher in the grade 2 sample than in the grade 1 sample, probably because a higher percentage of grade 2 students solved the items correctly (six of the seven items were identical to those in grade 1). Future work in this area may

be well advised to include some items at grade 2 that are slightly more difficult for students to solve correctly (i.e., that a lower percentage of grade 2 students will solve correctly).

The present report provides information about the items used, the data-collection procedure, and initial descriptive statistics. The 2015 sample involved a fairly large sample of teachers (*n* = 200) and students (*n* = 787). The data also included much more information about individual students' characteristics, such as performance on standardized tests, gender, race and ethnicity, and exceptionality. It also included information about teacher knowledge and beliefs, years of experience, educational background, and more. The size and scope of the set of data create many options for research on the topic of teacher judgment accuracy that uses the existing data.

Using a percent-correct approach based a fixed number of items in the denominator (i.e., 4, 7) would bias the measurement downward. We attempted to negate that bias, but because we did not account for potential differences in difficulty to predict student performance for the various missing items within some students, some bias remains in the result. We regard this threat of bias to be small, because the 2015 sample included only 12 cases of missing data (out of the 7 x 785 = 5495 data points), representing only two-tenths of 1% of the data.

4.1. Future Directions for Analysis and Inquiry

Very little is currently known about how teacher judgment accuracy is related to other factors such as student learning and teacher knowledge of more general principles such as mathematics content or teachers' instructional practice. We anticipate conducting future studies to investigate measurement techniques based on item response theory, associations between teacher judgment accuracy and student academic learning, the effect of the CGI intervention on teacher judgment accuracy, and potential bias in teacher judgment accuracy based on student characteristics.

Teachers and students in the treatment condition for the present study were randomly assigned to the CGI condition. A central aim of the CGI program is to focus teachers' attention on individual students' cognitive processes so that the teachers can use this information to shape their instructional decisions. A reasonable hypothesis, therefore, is that the teachers in the schools assigned to the treatment condition have higher levels of teacher judgment accuracy than their counterparts in the comparison-group schools. Descriptive statistics presented in Tables 20 and 21 of the present report indicate that the overall percentage of teacher judgment accuracy is higher in the treatment (CGI) group than the control group, but the present report does not represent a statistical test of the effect of the CGI program. Statistical models testing the impact of the CGI intervention on teacher judgment accuracy will need to be specified for investigation of this finding. Among other considerations, a test of the effect of the treatment condition may need to account for the nested structure of the data (i.e., students nested in teachers, teachers nested in schools).

Another important line of inquiry to be explored with these data will be the associations between teacher judgment accuracy and student learning. Teachers' knowledge of their individual students—as measured by teacher judgment accuracy—may mediate the effect of that teacher on student learning. Moreover, because formative assessment is believed to be an important factor in teaching and learning, and the CGI program is hypothesized to increase teacher judgment accuracy, the hypothesis seems reasonable that teacher judgment accuracy mediates the effect of the CGI program on student achievement. Other correlates to explore may be teachers' general levels of mathematical knowledge for teaching at the early elementary level, teachers' beliefs about mathematics teaching and learning, and teachers' instructional practice.

One of the more important investigations made possible by the current data set is of potential bias in teacher judgment accuracy based on student characteristics. The elements of underprediction and overprediction will provide opportunities to examine the possible correlations between teacher judgment accuracy and student characteristics such as gender, exceptionality, or English-language-learner status. Measures of teacher judgment accuracy may therefore provide insight into critically important matters of equity in mathematics.

4.2. Summary and Conclusions

In summary, we think the TAPSS instrument represents an important development in the field of teacher judgment accuracy. The design of the instrument involves the highest levels of specificity, congruence, and informed judgment that have been identified in the research literature. Moreover, the sample size of the two field-tests of the TAPSS instrument is higher than those of most published research. Rather than strictly focusing on deficiencies in teacher knowledge or expectations about students, the TAPSS instrument has the capacity to measure accurate and inaccurate inferences that teachers make about students. As a result, the data from the two field-tests of the TAPSS instrument have high potential for helping to advance the field and make an important contribution in this particular area.

We do not currently know how teacher judgment accuracy is related to other factors such as student learning, teacher knowledge of more generalized principles such as mathematics content or student learning progressions, or teachers' instructional practice. Future analyses of these data will investigate measurement techniques based on multilevel modeling that takes into account group effects, scoring techniques based on item response theory, associations between teacher judgment accuracy and student academic learning, the effect of the CGI intervention on teacher judgment accuracy, and potential bias in teacher judgment accuracy based on student characteristics. Ultimately, a better theoretical understanding is needed for analysis of the three aspects of teachers' ability to assess student success and how they influence each other.

For many decades, researchers and educational leaders have sought to identify malleable facets of teacher knowledge that can predict student learning and increase student academic achievement. The search has proven quite challenging. Teacher judgment accuracy may ultimately serve to provide an important missing link in the study of this important topic.

References

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389–407.

Campbell, P. F., Rust, A. H., Nishio, M., DePiper, J. N., Smith, T. M., Frank, T. J., Clark, L. M., Griffin, M. J., Conant, D. L., & Choi, Y. (2014). The relationship between teachers' mathematical content and pedagogical knowledge, teachers' perceptions, and student achievement. *Journal for Research in Mathematics Education, 45*(4), 419–459.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (2015). *Children's Mathematics: Cognitively Guided Instruction.* (2nd ed.) Portsmouth, NH: Heinemann.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C.-P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal, 26*(4), 499–531.

de Boer, H., Bosker, R. J., & van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology, 102*(1), 168–179.

Dunbar, S. B., Hoover, H. D., Frisbie, D. A., Ordman, V. L., Oberley, K. R., Naylor, R. J., & Bray, G. B. (2008). *Iowa Test of Basic Skills®, Form C, Level 7*. Rolling Meadows, IL: Riverside Publishing.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education, 27*(4), 458–477.

Carpenter, T. P., & Franke, M. L. (2004). Cognitively Guided Instruction: Challenging the core of educational practice. In T. K. Glennan, S. J. Bodilly, J. R. Galegher & K. A. Kerr (Eds.), *Expanding the reach of education reforms: Perspectives from leaders in the scale-up of educational interventions* (pp. 41-80). Santa Monica, CA: RAND Corporation.

Gabriele, A. J., Joram, E., & Park, K. H. (2016). Elementary mathematics teachers' judgment accuracy and calibration accuracy: Do they predict students' mathematics achievement outcomes? *Learning and Instruction, 45*, 49–60.

Harvey, K. E., Suizzo, M.-A. & Jackson, K. M. (2016). Predicting the grades of low-income–ethnic-minority students from teacher-student discrepancies in reported motivation. *The Journal of Experimental Education, 84*(3), 510–528.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371–406.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297–313.

Jussim, L., & Eccles, J. (1992). Teacher expectations. II: Construction and reflection of student achievement. *Journal of Personality and Social Psychology, 63*(6), 947–961.

Lang, L. B., Schoen, R. C., LaVenia, M., & Oberlin, M. (2014). *Mathematics Formative Assessment System–Common Core State Standards: A randomized field trial in kindergarten and first grade.* Paper presented at the annual spring conference of the Society for Research in Educational Effectiveness, Washington, DC.

Lang, L. B., Schoen, R. C., LaVenia, M., Oberlin, M., & Robinson, M. (2013). *K–3 mathematics formative assessment: Effects on teaching, learning, and the gender gap.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Madon, S., Jussim, L., & Eccles, J. (1997). In search of the powerful self-fulfilling prophecy. *Journal of Personality and Social Psychology, 72*(4), 791–809.

NGACBP (National Governors Association Center for Best Practices) & CCSSO (Council of Chief State School Officers) (2010). *Common Core State Standards for Mathematics*. Washington, DC: Author.

Ready, D., & Wright, D. (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.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development.* New York: Rinehart and Winston.

Sarama, J. & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children.* New York: Routledge.

Schoen, R. C., Bray, W., Wolfe, C., Tazaz, A. M., & Nielsen, L. (2017). Developing an assessment instrument to measure early elementary teachers' mathematical knowledge for teaching. *The Elementary School Journal, 118*(1), 55–81.

Schoen, R. C., LaVenia, M., Bauduin, C., & Farina, K. (2016a). *Elementary mathematics student assessment: Measuring the performance of grade 1 and 2 students in counting, word problems, and computation in fall 2013* (Research Report No. 2016-03). Tallahassee, FL: Learning Systems Institute, Florida State University.

Schoen, R. C., LaVenia, M., Bauduin, C., & Farina, K. (2016b). *Elementary mathematics student assessment: Measuring the performance of grade 1 and 2 students in counting, word problems, and computation in fall 2014* (Research Report No. 2016-04.) Tallahassee, FL: Learning Systems Institute, Florida State University.

Schoen, R. C., LaVenia, M., Champagne, Z. M., & Farina, K. (2016). *Mathematics performance and cognition (MPAC) interview: Measuring first- and second-grade student achievement in number, operations, and equality in spring 2014* (Research Report No. 2016–01). Tallahassee, FL: Learning Systems Institute.

Schoen, R. C., LaVenia, M., Champagne, Z. M., Farina, K., & Tazaz, A. (2016). *Mathematics Performance and Cognition (MPAC) interview: Measuring first- and second-grade student achievement in number, operations, and equality in spring 2015* (Research Report No. 2016-02.) Tallahassee, FL: Learning Systems Institute, Florida State University.

Schoen, R. C., LaVenia, M., Tazaz, A., & Farina, K. (2017a). *Replicating the CGI experiment in diverse environments: Effects of year 1 on student mathematics achievement* (Research Report No. 2017–01.) Tallahassee, FL: Learning Systems Institute, Florida State University.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Südkamp, A., Kaiser, J., & Möller , J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762.

Appendix A – Teacher and Student Sample Demographic Tables

Table 22. 2013–14 Teacher Sample Demographics

Characteristics	Sample by condition				Sample by grade				Predicting analytic sample	
	Treatment (*n* = 67)		Control (*n* = 79)		Grade 1 (*n* = 77)		Grade 2 (*n* = 69)		(*n* = 146)	
	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*
Grade										
1	.51	34	.54	43	—	—	—	—	.53	77
2	.49	33	.46	36	—	—	—	—	.47	69
Condition										
Treatment	—	—	—	—	.44	34	.48	33	.46	67
Control	—	—	—	—	.56	43	.52	36	.54	79
Years of teacher experience										
Three or fewer	.24	16	.15	12	.23	18	.14	10	.19	28
Four or more	.76	51	.85	67	.77	59	.86	59	.81	118
Gender										
Female	.99	66	1.0	79	.99	76	1.0	69	.99	145
Male	.01	1	.00	0	.01	1	.00	0	.01	1
Race/Ethnicity										
Asian/Pacific Islander	.00	0	.01	1	.00	0	.01	1	.01	1
Black	.12	8	.08	6	.05	4	.15	10	.10	14
White	.76	51	.76	60	.81	62	.71	49	.76	111
Hispanic	.08	5	.14	11	.12	9	.10	7	.11	16

Note. No demographic information was missing for teachers in the analytic sample.

Table 23. 2014–15 Teacher Sample Demographics

Characteristics	Sample by condition				Sample by grade				Predicting analytic sample	
	Treatment (n = 85)		Control (n = 115)		Grade 1 (n = 106)		Grade 2 (n = 94)		(n=200)	
	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*
Grade										
1	.49	42	.56	64	—	—	—	—	.53	106
2	.51	43	.44	51	—	—	—	—	.47	94
Treatment Condition										
Treatment	—	—	—	—	.40	42	.46	43	.43	85
Control	—	—	—	—	.60	64	.54	51	.58	115
Years of teacher experience										
Three or fewer	.27	23	.21	24	.24	25	.23	22	.24	47
Four or more	.68	58	.74	85	.72	75	.72	68	.72	143
Missing	.05	4	.05	6	.06	6	.04	4	.05	10
Gender										
Female	.94	80	.92	106	.93	99	.93	87	.93	186
Male	.01	1	.03	3	.01	1	.03	3	.02	4
Missing	.05	4	.05	6	.06	6	.04	4	.05	10
Race/Ethnicity										
Asian/Pacific Islander	.01	1	0	0	0	0	.01	1	.01	1
Black	.13	11	.11	13	.09	10	.15	14	.12	24
White	.73	62	.67	77	.73	77	.66	62	.70	139
Hispanic	.08	7	.17	19	.12	13	.14	13	.13	26
Missing	.05	4	.05	6	.06	6	.04	4	.05	10

Note. Teachers with unreported demographic information are represented in the "Missing" category.

Table 24. 2013–14 Student Sample Demographics

Characteristics	Sample by condition				Sample by grade				Predicted analytic sample	
	Treatment (*n* = 227)		Control (*n* = 277)		Grade 1 (*n* = 277)		Grade 2 (*n* = 227)		(*n* = 504)	
	%	n	%	n	%	n	%	n	%	n
Grade										
1	.54	122	.56	155	—	—	—	—	.55	277
2	.46	105	.44	122	—	—	—	—	.45	227
Condition										
Treatment	—	—	—	—	.44	122	.46	105	.45	227
Control	—	—	—	—	.56	155	.54	122	.55	277
Gender										
Male	.49	112	.49	135	.50	138	.48	109	.49	247
Female	.50	115	.51	142	.50	139	.52	118	.51	257
Race/ethnicity										
Asian	.07	16	.04	11	.05	14	.06	13	.05	27
Black	.19	43	.20	55	.22	62	.16	36	.19	98
White	.40	92	.25	71	.27	76	.39	87	.32	163
Hispanic	.29	65	.49	135	.43	120	.35	80	.40	200
Other	.04	9	.01	4	.02	5	.03	8	.03	13
Missing	.01	2	.00	1	.00	0	.01	3	.01	3
English language learners										
ELL status	.18	40	.29	81	.24	67	.24	54	.24	121
Non ELL	.81	185	.70	195	.76	210	.75	170	.75	380
Missing	.01	2	.00	1	.00	0	.01	3	.01	3
Free/reduced lunch										
Eligible	.48	110	.74	205	.70	193	.54	122	.62	315
Not Eligible	.51	115	.26	71	.30	84	.45	102	.37	186
Missing	.01	2	.00	1	.00	0	.01	3	.01	3
Exceptionality										
Students with disabilities	.06	13	.07	18	.07	19	.05	12	.06	31
Gifted	.09	20	.03	9	.04	11	.08	18	.06	29
Missing	.09	2	.00	1	.00	0	.01	3	.01	3

Note. Students with unreported demographic information are represented in the "Missing" category. The Race/ethnicity categories are mutually exclusive. The Exceptionality categories are not mutually exclusive.

Table 25. 2014-15 Student Sample Demographics

Characteristics	Sample by condition				Sample by grade				Predicted analytic sample	
	Treatment (*n* =334)		Control (*n* = 451)		Grade 1 (*n* = 420)		Grade 2 (*n* = 365)		(*n* = 785)	
	%	n	%	n	%	n	%	n	%	n
Grade										
1	.50	168	.56	252	—	—	—	—	.54	420
2	.50	166	.44	199	—	—	—	—	.46	365
Condition										
Treatment	—	—	—	—	.40	168	.46	168	.43	336
Control	—	—	—	—	.60	252	.54	199	.57	451
Gender										
Male	.49	165	.50	226	.50	210	.50	181	.50	391
Female	.51	171	.50	225	.50	210	.50	184	.50	394
Race/ethnicity										
Asian	.07	22	.06	26	.06	25	.06	23	.06	48
Black	.19	62	.15	67	.15	61	.19	68	.16	129
White	.39	130	.30	135	.32	133	.36	132	.34	265
Hispanic	.23	78	.38	170	.35	146	.28	102	.32	248
Other	.03	11	.02	10	.03	11	.03	10	.03	21
Missing	.09	31	.10	43	.10	44	.08	30	.09	74
English language learners										
ELL status	.17	58	.29	130	.25	107	.22	81	.24	188
Non ELL	.73	245	.62	278	.64	269	.70	254	.67	523
Missing	.09	31	.10	43	.10	44	.08	30	.09	74
Free/reduced lunch										
Eligible	.46	153	.62	280	.58	242	.52	191	.55	433
Not Eligible	.45	150	.28	128	.32	134	.39	144	.35	278
Missing	.09	31	.10	43	.10	44	.08	30	.09	74
Exceptionality										
Students with disabilities	.06	20	.06	26	.06	24	.06	22	.06	46
Gifted	.03	9	.02	11	.01	5	.04	15	.03	20
Missing	.09	31	.10	43	.10	44	.08	30	.09	74

Note. Students with unreported demographic information are represented in the "Missing" category. The Race/ethnicity categories are mutually exclusive. The Exceptionality categories are not mutually exclusive.

Appendix B – Teacher Prediction Sheets

Spring 2014 Teacher Prediction Sheet

Teacher _____ Date _____

School _____ Grade Level _____

Below is a list of four children selected at random from your classroom. For each problem listed on the attached page, predict whether or not you would expect that child to solve the given problem correctly. The students will have paper, markers, and manipulatives available to use.

For the first two problems, simply write YES or NO to indicate whether you think the child will correctly solve the equation for the missing number.

For the three basic facts problems, write whether you think child will (A) generate a correct answer and (B) know the sum or difference of those two numbers at a recall level. In the first column, write YES if you think the child will correctly answer or NO if you think the child will not correctly solve the problem. In the second column provided for each basic fact problem, write YES or NO to indicate whether the child knows the fact at a recall level. If you think the child has the fact memorized, write YES in the second column for each problem. If you think the child does not have the fact memorized, write NO in the second column. (If you do not think the child will produce a correct answer, you can skip the recall part of the question.)

Student	5 + ☐ = 13	6 + 3 = ☐ + 4	6 + 5 = ☐		12 – 7 = ☐		4 + 8 = ☐	
	Correct (Yes/No)	Correct (Yes/No)	Correct (Yes/ No)	Recall (Yes/ No)	Correct (Yes/ No)	Recall (Yes/ No)	Correct (Yes/ No)	Recall (Yes/ No)

Spring 2015 Grade One Teacher Prediction Sheet

Teacher _____ Date _____

School _____ Grade Level _____

Below is a list of four children selected at random from your classroom. For each problem listed, predict whether or not you think the child will solve the given problem correctly. For the True/Not True and missing addend item, the students will not have access to manipulatives or pencil/paper. The students will be asked to solve them mentally and/or with their fingers. The students will have paper, a marker, snap cubes, and base ten blocks available to use for the two subtraction computation problems and the two word problems.

For the True/Not True questions, write the response (True or Not True) you think the student will give. For the multidigit subtraction problems, word problems, and missing addend problem, write YES if you think the child will provide the correct answer or NO if you think the child will not provide the correct answer.

	10 = 7 + 3	6 = 6	6 + 3 = ☐ + 4	102 – 3	21 – 19	James worked on his homework for 8 minutes. Courtney worked on her homework for 15 minutes. How many minutes longer did Courtney work on her homework than James?	Caleb had 15 books on his shelf. Then, he got some more books from the library and put them on his shelf. Now, he has 24 books on his shelf. How many books did Caleb get from the library?
Student	True or Not True	True or Not True	Correct (Yes/No)	Correct (Yes/No)	Correct (Yes/No)	Correct (Yes/No)	Correct (Yes/No)

Spring 2015 Grade Two Teacher Prediction Sheet

Teacher _____ Date _____

School _____ Grade Level _____

Below is a list of four children selected at random from your classroom. For each problem listed, predict whether or not you think the child will solve the given problem correctly. For the True/Not True and missing addend item, the students will not have access to manipulatives or pencil/paper. The students will be asked to solve them mentally and/or with their fingers. The students will have paper, a marker, snap cubes, and base ten blocks available to use for the two subtraction computation problems and the two word problems.

For the True/Not True questions, write the response (True or Not True) you think the student will give. For the multidigit subtraction problems, word problems, and missing addend problem, write YES if you think the child will provide the correct answer or NO if you think the child will not provide the correct answer.

	10 = 7 + 3	6 = 6	6 + 3 = ☐ + 4	102 – 3	21 – 19	James worked on his homework for 8 minutes. Courtney worked on her homework for 15 minutes. How many minutes longer did Courtney work on her homework than James?	Aiden has collected 25 cards. He wants to collect 44 cards in total. How many more cards does Aiden need to collect?
Student	True or Not True	True or Not True	Correct (Yes/No)	Correct (Yes/No)	Correct (Yes/No)	Correct (Yes/No)	Correct (Yes/No)