The Validation of a Student Survey on Teacher Practice

Ryan Balch

Vanderbilt University

Abstract: Though there is evidence that teacher quality matters, a more challenging problem exists in attempting to measure teacher effectiveness. The current paper outlines the validation process for a student survey on teacher practice.
“It can be argued that student opinions of a teacher are an important consideration in any teacher evaluation system because students have the most contact with the teachers and are the direct consumers of a teacher’s service” (Goe, Bell, & Little, Approaches to Evaluating Teacher Effectiveness: A Research Synthesis, 2008, p. 39).

As we struggle to develop measures of teacher effectiveness, the statement above reflects the belief that students represent a potentially valuable source of information in evaluating a teacher’s performance. Though more than one-third of U.S. companies use some type of multi-source feedback to evaluate employees, education has been slow to solicit feedback from anyone but superiors (Bracken, Timmereck, & Church, 2001a). In an era of increased accountability for both teachers and schools, our determinations of effective teaching are only as valid as the instruments of evaluation.

Though we still seek an agreed upon definition of effective teaching, one consistent finding in education research is that there is significant variation among teachers in their ability to increase student achievement. Hanushek & Rivkin (2006) find that teachers near the top end of the quality distribution can lead their students to a full year’s worth of learning more than teachers near the bottom end. Specifically, teachers in the 95th percentile had students gain 1.5 grade level equivalents while teachers in the 5th percentile only had an increase of 0.5 during one school year. Even using a more conservative estimate, others have found that moving from an average teacher to one in the 85th percentile can result in a 4 percentile average test score gain for students in that teacher’s class (Rivkin, Hanushek, & Kain, 2005). In the same data, this was roughly equivalent to reducing class size by 10 students.

Further research has demonstrated that these substantial differences in achievement are both additive and cumulative for students. Having an effective teacher for three sequential years resulted in a 50 percentile point difference in achievement compared to students who experienced an ineffective teacher for three years (Sanders & Rivers, 1996). Moreover, subsequent teachers also appear unable to reverse either the positive or negative effects of previous teachers.

While it is clear that there are differences among teachers, it is still challenging to quantify and measure this variation. Researchers and practitioners have sought to meet this challenge using several different methods of teacher evaluation ranging from observational evaluation rubrics and teacher portfolios to value-added calculations for student achievement. A less common method uses feedback from students to measure teacher quality, though there is no instrument that has been routinely employed in schools. The goal of the current investigation is to develop a valid and reliable instrument for measuring teacher effectiveness using student feedback. To assist in incorporating the student perspective, the current investigation outlines the development of a valid and reliable instrument to measure teacher effectiveness using student feedback.

The current student feedback survey was developed using commonalities among established observational rubrics such as Danielson’s (1996) framework for teaching and a
thorough literature review of teacher behaviors that are found to consistently predict student achievement (Marzano, Pickering, & Pollock, 2001; Brophy & Good, 1986; Pianta, Paro, & Hamre, 2006; Schacter & Thum, 2004; Emmer & Evertson, 1981). The overall categories include presentation style, lesson structure, behavior management, productivity, teacher-student relationships, awareness of student need, feedback, challenge, engaging and motivating students, as well as content expertise.

Student feedback on teacher practice is valuable for several reasons. First, it offers a complement to value-added measures of student achievement as a measure of teacher quality. Though academic learning is one thing that we may value in education, student achievement does not measure how well teachers perform in other areas such as developing engagement and confidence in learning. Second, value-added is not feasible for many teachers because they teach subjects that do not have standardized tests such as electives or some high school subjects. Without secondary measures, these teachers may end up being judged by the performance of the school rather receiving any individual feedback. Third, when multiple teachers teach students, such as reading teachers or special education team teachers, it is difficult to attribute achievement gains to a particular teacher. With surveys, students can be directed to answer questions about the practices of a specific teacher. Fourth, it provides reliable and valid information on specific practices rather than solely overall effectiveness. Student surveys can therefore form the basis of targeted professional development in the areas where teachers obtain low ratings. Fifth, in comparison to observational evaluations, student surveys provide similar information at a fraction of the cost and time for implementation. Finally, using multiple measures in tandem can prevent manipulation and gaming behavior in a high stakes environment.

Following the process outlined by Porter et al. (2010), the current survey will undergo several rounds of validation to ensure that survey questions adequately assess various dimensions of teacher quality. This will include multiple validation checks with teachers and content experts as well as two rounds of cognitive interviews with high school students as a means of aligning student understanding of survey questions with the intended question objective. The final validation step includes two rounds of pilot testing to assess internal reliability and criterion validity. The first round was conducted with 140 students at a local high school in order to assess internal reliability. The second round is to be conducted in the spring of 2011. Teachers’ total score on the student feedback survey will be compared to their mean value-added student achievement, student engagement, and student self-efficacy to assess the predictive validity of scales and individual items. Overall, the investigation will be guided by the following research questions:

- What is the relationship between a teacher’s total score and scale scores on a student survey and their average gains in student achievement?
- What is the relationship between a teacher’s total score and scale scores on a student survey and scores on instruments measuring student engagement and student self-efficacy?
The next section reviews the literature on teacher quality and teacher evaluation. Following this is a description of the survey development process and an outline of the validation process.

**Review of Literature**

**Teacher Quality**

A precursor to developing measures of teacher quality is agreeing upon a definition. Complicating the matter, teacher quality can be defined in a number of ways. These may include teacher qualifications, teacher characteristics, or actual teaching behaviors, with each showing a different relationship to student achievement.

Teacher qualifications include aspects such as teaching experience, advanced degrees, certification, and subject matter knowledge. Teacher experience predicts student achievement in some studies (Clotfelter, Ladd, & Vigdor, 2006; Harris & Sass, 2007), but often the effect is limited to the first few years of a teacher’s career (Hanushek, Kain, O’Brien, & Rivkin, 2005; Rockoff J., 2004). Other studies find no relationship between experience and student achievement (Jacob & Lefgren, 2005). For level of education, research consistently fails to find a relationship between advanced degrees and student achievement (Harris & Sass, 2007; Hanushek, Kain, O’Brien, & Rivkin, 2005; Clotfelter, Ladd, & Vigdor, 2006). Overall, Goldhaber (2002) finds that only 3 percent of a teacher’s contribution to student learning was associated with teacher experience, degree attained, or other observable characteristics. These results call into question the fact that the vast majority of district salary structures reward teachers for qualifications – advanced degrees and years of teaching experience – that bear little relationship to student outcomes.

A more promising teacher qualification is teacher content knowledge. Many of the studies investigating this aspect have sought to use teacher certification scores as a proxy for content knowledge, with the results generally showing a positive relationship (Greenwald, Hedges, & Laine, 1996; Rowen, Correnti, & Miller, 2002; Ferguson, 1991). These measures, however, are general measures of content that do not inform the types of knowledge or ability that a teacher requires to be an effective teacher. Other measures have looked specifically at performance on instruments designed to test a teacher’s mathematical knowledge for teaching and found a significant and positive relationship to student achievement (Hill, Rowan, & Ball, 2005). Although there appears to be evidence of a link between content knowledge and achievement, the type of content knowledge that is assessed is dependent on the instrument or measure.

Next, a large body of research has investigated what teacher characteristics are most associated with increased student achievement, with no clear consensus that any measured characteristics have an impact (Goe, 2007; Rockoff, Jacob, Kane, & Staiger, 2008; Goldhaber, 2002). Characteristics such as race, ethnicity, and gender do not have a significant relationship
to student achievement (Ehrenberg, Goldhaber, & Brewer, 1995), but there is evidence from the STAR randomized class size experiment that students with teachers of the same race have increased achievement in both reading and math (Dee, 2004). Rockoff et al. (2008) investigated a range of non-traditional teacher characteristics including content knowledge, cognitive ability, personality traits, and feelings of self-efficacy. They find that very few of these predictors have a strong relationship to achievement when analyzed individually, but factors that combine cognitive and non-cognitive teacher skills have a modest relationship (Rockoff, Jacob, Kane, & Staiger, 2008).

Finally, teacher quality may be defined by actual teacher behaviors that are associated with increased student achievement. Beginning in the 1960’s and 1970’s, there was a push to determine what teacher practices were associated with increased student achievement (Schacter & Thum, 2004). A number of reviews have consolidated findings from these individual studies in an effort to present behaviors that show consistent relationships to student achievement. The categories from major reviews are shown in Table 1 below.

While there are differences, a considerable amount of overlap exists among these reviews. For instance, providing high quality academic feedback is noted in several reviews as having an association with higher student achievement. Schachter and Thum (2004) call for “frequent, elaborate, and high quality academic feedback”, Good and Brophy (1986) note the importance of “monitoring students’ understanding, providing feedback, and giving praise”, Emmer and Evertson (1994) emphasize that “all student work, including seatwork, homework, and papers, is corrected, errors are discussed, and feedback is provided promptly”, and Marzano (2001) outlines a large body of research indicating the importance of teacher providing feedback. Other categories that overlap among many of the reviews include clarity of presentation, managing behavior promptly, reinforcing student effort, and having appropriate pacing.

Table 1 - Literature Reviews of Teaching Behaviors

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Questions – Type, frequency, required student response, wait time</td>
<td>Clarity about instructional goals</td>
<td>Clarity about instructional goals</td>
<td>Providing students with the opportunity to learn what is to be tested</td>
<td>Rules and Procedures – established and enforced and students are monitored for compliance</td>
<td>Identifying similarities and differences</td>
<td></td>
</tr>
<tr>
<td>Grouping – strategies for cooperative learning</td>
<td>Thinking – Metacognition generative learning</td>
<td>Knowledge of content and ways for teaching it</td>
<td>Variety in the use of teacher methods and media</td>
<td>Realistic praise – not praise for its own sake</td>
<td>Consistency – Similar expectations are maintained by activities and behavior at all times for all students</td>
<td></td>
</tr>
</tbody>
</table>
| Feedback – Frequent, elaborate, and high quality academic feedback | Feedback – Meaningful projects and simulations to foster opportunities for learning by doing and | Making comments that help structure learning of knowledge and concepts for students, helping students learn how to learn | Prompt Management of inappropriate behavior | Academic instruction – Attention is focused on the | Re...
### Measures of Teacher Behaviors

While the knowledge of what teacher qualifications or behaviors may be associated with student achievement is important, this knowledge is only useful if it can be measured. This is especially true when one considers information asymmetry from the principal-agent framework. Principal-agent theory describes the situation in which an employer (principal) hires an employee (agent) to perform a task (Alchian & Demsetz, 1972). The problem arises when the agent receives the same compensation regardless of the quality of work or effort level, sometimes leading to a reduction in both (Eisenhardt, 1989). The principal-agent problem is especially relevant in situations where there is not a clearly defined output of performance and low levels of supervisor monitoring, situations that occur frequently in the teaching profession. When employees have lower incentives for increasing effort, it is argued that it is more efficient to replace fixed wages with compensation that links an employee’s pay to their performance (Alchian & Demsetz, 1972). In this regard, the type of incentives and method of measurement impact ways that systems address the inherent problem that principal-agent theory outlines.

<table>
<thead>
<tr>
<th>Task Orientation or businesslike teacher behaviors, structures, routines, and academic focus</th>
<th>student interaction</th>
<th>management of student work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesson Objectives- Objectives explicitly communicated</td>
<td>Motivating students – Attend to students notions of competence, reinforcing student effort</td>
<td>With-it-ness – awareness of what is going on, alertness in monitoring classroom activities Overlapping – sustaining an activity while doing something else at the same time</td>
</tr>
<tr>
<td>Checking student work – All student work, including seatwork, homework, and papers, is corrected, errors are discussed, and feedback is provided promptly</td>
<td>Homework and Practice</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Student Opportunity to Learn, that is, the teacher’s coverage of the material or content in class on which students are later tested</th>
<th>Presentation – Illustrations, analogies, modeling by teacher, concise communication</th>
<th>Teacher Knowledge of Students – prior knowledge, incorporating student interest through differentiated approaches</th>
<th>Monitoring of students’ understanding, providing appropriate feedback, giving praise, asking questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interaction teaching – Presenting and explaining new material, question sessions, discussions, checking for student understanding, actively moving among students, and providing feedback</td>
<td>Nonlinguistic representations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>“Promising” -Using student ideas -Justified criticism -Using structuring comments</th>
<th>Classroom Environment – Student discipline and behavior, student work ethic, teacher caring for individual pupils</th>
<th>Smoothness – Sustaining proper lesson pacing and group momentum, not dwelling on minor points or wasting time dealing with individuals, and focusing on all students Flexibility in planning and adapting classroom activities</th>
<th>Clarity – Lessons are presented logically and sequentially. Clarity is enhanced by the use of instructional objectives and adequate illustrations and by keeping in touch with students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperation Learning Questions, Cue, and advance organizers</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>-Encouraging student elaboration -Using challenging instructional materials</th>
<th>Seatwork instructions and management that initiate and focus on productive task engagement</th>
<th>Pacing – Information is presented at a rate appropriate to the students’ ability to comprehend it</th>
<th>Setting Objectives and Providing Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generating and testing hypothesis</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| -Asking appropriate questions suited to students’ cognitive level | Holding students accountable for learning; accepting responsibility for student learning | Transitions – Transitions from one activity to another are made rapidly, with minimal confusion | Generating and testing hypothesis |
There are a variety of instruments and techniques that have been implemented to measure teacher behaviors that will now be discussed.

**Self-Evaluation**

Two techniques that require teachers to assess their own practices are teacher surveys and logs. Many national surveys ask teachers about practices used during the entire year such as the Early Childhood Longitudinal Study, the National Assessment of Educational Progress, or the Schools and Staffing Survey’s Teacher Follow-Up Survey. While these surveys tap a nationally representative population, they require that teachers make assessments of their practice from the entire year and may be subject to teachers responding with socially desirable answers or error due to problems with remembering (Rowan, Jacob, & Correnti, 2009). An alternative to large surveys is the use of instructional logs, a process by which teacher document content coverage and teaching strategies on a more regular basis. While not as accurate as independent observation because of bias due to self-report, instructional logs have been found to be valid, reliable, and cost-effective (Rowan, Jacob, & Correnti, 2009). To establish validity, teacher logs were compared to researcher evaluation, finding that teacher-observer match rates ranged from 73 to 90 percent. Though possible for documentation of instructional practice, it is unlikely that this rate would persist in a high-stakes environment.

**Analysis of Classroom Artifacts and Portfolios**

A second possible method for evaluation includes the analysis of classroom artifacts such as lesson plans, teacher assignments, assessments, scoring rubrics, and student work. While many systems use some sort of artifact analysis, a structured and valid protocol for evaluation is essential. Examples of such protocols include the Instructional Quality Assessment done by the National Center for Research on Evaluation, Standards, and Student Testing (Matsumura, Slater, Junker, et al., 2006) and the Intellectual Demand Assignment Protocol (IDAP) developed by the Consortium on Chicago School Research (Newmann et al., 2001). The IDAP showed both high interrater reliability (90 percent agreement) and that students of teachers that scored high on the instrument had learning gains on the Iowa Test of Basic Skills that were 20 percent higher than the national average. Though some findings indicate that teacher ratings using these artifacts are correlated with outcomes, there is a lack of research conducted by independent researchers (Goe, Bell, & Little, Approaches to Evaluating Teacher Effectiveness: A Research Synthesis, 2008).

Portfolios are a further option that may include similar teaching artifacts, yet in this case teachers prepare their own samples. It also may include other evidence such as statements of teaching philosophy or videotaped lessons. A common example of this is National Board for Professional Teaching Standards Certification, which research indicates is successful in identifying high-quality teachers even though the process of certification may not improve.
effectiveness (Hakel, Koenig, & Elliot, 2008). It also remains to be seen whether this type of evaluation would be practical in a high-stakes setting.

**Classroom Observation**

Classroom observations represent one of the most commonly used evaluation systems for teachers (Goe, 2008). There are countless variations in frequency, instrument, rating scales, and protocol. Some of the main issues to consider with observations are the validity of the instrument and the reliability of rating, particularly if ratings are attached to financial rewards or job security. Examples of instruments that have been validated for their relationship to student achievement include Charlotte Danielson’s (1996) *Framework for Teaching* and the Classroom Assessment Scoring System (CLASS) for grades K-5 (Pianta, La Paro, & Hamre, 2006). Outside researchers found that student with a teacher in the top quartile according to Danielson’s rubric would score 0.10 standard deviations higher in math and 0.125 standard deviations higher in reading than a student assigned to a teacher in the bottom quartile (Kane, Taylor, Tyler, & Wooten, 2010).

Observational rubrics consolidate research on teacher behaviors that are associated with increased student achievement and provide hierarchical levels for observers to use as a basis for judgment. For example, Schacter and Thum (2004) developed six teaching standards of teacher behavior in the areas of questions, feedback, presentation, lesson structure and pacing, lesson objectives, and classroom environment. The categories were based on teaching models that combined effective practices and garnered moderate to large effect sizes (d = 0.46 – 1.53) in reading, language, mathematics, and social science (Gage & Needles, 1989).

These were combined with five standards of teaching strategies found to show increased student achievement that included grouping students, encouraging student thinking, providing meaningful activities, motivating students, and teacher knowledge of students. All together Schacter and Thum developed a rubric with 12 different teaching standards and a corresponding rubric to determine teacher quality. The rubric was tested with 52 elementary school teachers and the authors found that students of teachers who implement the practices in these 12 categories make considerable gains in achievement (Schacter & Thum, 2004).

While observations using these rubrics have demonstrated a link to student achievement, the investment necessary for complete implementation is large. Extensive training is necessary for all evaluators, and immediate connection to incentives may increase the potential for errors due to inexperience with an evaluation instrument. Though observational evaluation is the most common form of personnel evaluation for teachers, the rubrics many school systems employ do not require the training or expertise necessary for more advanced instruments.

**Student Feedback**
The majority of empirical investigations of student feedback in teacher evaluation have occurred within higher education. From these studies, student ratings appear to be both reliable and valid measures of instructional quality. Ratings in subsequent years have correlations between .87 and .89, suggesting they are stable and reliable (Aleamoni, 1999). Further, a meta-analysis of validity studies on student ratings found an average correlation of .43 between mean student ratings of instructors and mean student performance on common final exams in multi-section courses. This is combined with positive correlations between student feedback and ratings from colleagues and external observers (Renaud & Murray, 2005). As Renaud and Murray (2005) note in their review of the literature, “the weight of evidence from research is that student ratings of teacher effectiveness validly reflect the skill or effectiveness of the instructor” (p. 930).

Despite these findings, it is possible that extraneous factors could bias these ratings. Some have found a negative relationship between student ratings and expected course grades, indicating that students rate challenging teachers lower (Rodin & Rodin, 1972). In an extensive review, it was found that twenty-four studies found zero relationship, and thirty-seven studies found contradictory results to this notion (Aleamoni, 1999). Clearly, students are able to separate effective instruction from their own personal academic expectations.

Further, an instructor’s personality characteristics could influence students’ perception of their effectiveness, resulting in positively biased ratings. A review of the literature finds seventeen studies that find students are “discriminating judges of instructional effectiveness,” (Aleamoni, 1999, p. 154) with students showing considerable ability to differentially rate instructors in categories such as method of instruction, course content, general course attitude, and interest and attention. While some overlap between an instructor’s personality characteristics and their effectiveness should be expected, it is important to document that student ratings can disentangle these concepts.

In contrast to higher education, the literature on student feedback in k-12 settings is less extensive. Though evidence stems from only four main investigations, the promising results suggest surveys have the potential to serve as an alternative measure of teacher effectiveness. First, in a study of 1976 K-12 students in Wyoming, Wilkerson et al. (2000) found that student ratings were highly significant predictors of student achievement in reading (p<.001) while self-ratings by teachers, principal ratings, and principal summative evaluations were not significant at even the .05 level. They further demonstrated that student ratings were significant predictors of both Language Arts and math. The correlations from this study are shown in Table 2.
Similarly, in a study of 9,765 student surveys, researchers found that student surveys at various levels (elementary, middle, and high school) were valid and reliable teacher evaluation measures (Peterson, Wahlquist, & Bone, 2000). This aligns with international research from Cyprus where student surveys of teacher practices were highly correlated with achievement gains in math and Greek language as well as other affective outcomes of schooling (Kyriakides, 2005). While more research is needed, these findings “provide convincing evidence that student ratings of teaching are worth considering for inclusion in teacher evaluation systems” (Goe, Bell, & Little, Approaches to Evaluating Teacher Effectiveness: A Research Synthesis, 2008, p. 40).

More recently, an extensive research project funded by the Gates Foundation is investigating the relationship between several measures of teaching and value-added student achievement. Referred to as the Measures of Effective Teacher (MET) Project, these measures include observational rubrics, tests of teaching strategies, and student perceptions (Kane & Cantrell, 2010). Student perceptions are measured using the 36 question Tripod student perception survey developed by Ron Ferguson at Harvard University. Preliminary findings from the project report a significant correlation in both math (.212) and ELA (.095) between a teacher’s total score and value-added achievement on state tests.

**Methods**

There are multiple steps involved in developing a psychometrically sound instrument. Following the process outlined by Porter et al. (2010), the current survey will undergo several rounds of validation and two stages of pilot testing to assess and improve the validity and reliability of the survey. These include establishing content validity, conducting cognitive interviews with students, and pilot testing to assess internal reliability and predictive validity. Each of these processes will be outlined below.

**Content Validity**

The main source of content validity comes from the survey development process, with questions developed using two main procedures. The first procedure consisted of identifying
overlapping teacher behaviors from the various reviews of the literature. For instance, all of the reviews highlight a link between providing feedback for students and higher student achievement. Schachter and Thum (2004) note that teachers should provide “frequent, elaborate, and high quality academic feedback”, Good and Brophy (1986) discuss “monitoring of students’ understanding and providing appropriate feedback”, Emmer and Evertson (1994) note that “all student work, including seatwork, homework, and papers, is corrected, errors are discussed, and feedback is promptly provided”, and finally Marzano (2001) outline several research based feedback strategies.

When a commonality among the reviews is found, the teacher behavior is then written into a behavior that can be observed. Schacter and Thum’s (2004) observational rubric definitions helped with this process when appropriate. For instance, the category of feedback contains several behaviors identified by the rubric for ‘Exemplary’ teachers. Table 3 displays some of the behaviors identified by the rubric and the corresponding survey questions.

Table 3 - Rubric Behavior and Corresponding Survey Question

<table>
<thead>
<tr>
<th>Schacter and Thum Rubric Behavior</th>
<th>Corresponding Student Survey Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback makes students explicitly aware of performance criteria in the form of rubrics or criterion charts.</td>
<td>My teacher gives us guidelines for assignments (rubrics, charts, grading rules, etc) so we know how we will be graded.</td>
</tr>
<tr>
<td>Teacher engages students in giving specific and high quality feedback to one another.</td>
<td>I have opportunities during this class to give and receive feedback from other students.</td>
</tr>
<tr>
<td>Feedback from students is used to monitor and adjust instruction.</td>
<td>My teacher changes the activity or lesson if many students do not understand.</td>
</tr>
<tr>
<td>The teacher circulates to prompt student thinking, assess each student’s progress, and provide individual feedback.</td>
<td>My teacher walks around the room to check on students when we are doing individual work in class</td>
</tr>
</tbody>
</table>

The second procedure involved using common observational rubrics such as Charlotte Danielson’s (1996) *Framework for Teaching* and the Classroom Assessment Scoring System (CLASS) for grades K-5 (Pianta, La Paro, & Hamre, 2006). Both of these instruments have been thoroughly tested for criterion validity by assessing the relationship between teacher scores on the rubric and a teacher’s value-added student achievement (Kane, Taylor, Tyler, & Wooten, 2010). These also represent the two rubrics chosen to measure general teaching practice in seven large school districts as part of the current Measures of Effective Teaching project sponsored by the Gates Foundation.

To further develop the survey, teacher behaviors identified by the highest levels of the rubric were transformed into questions appropriate for students to answer. There was considerable overlap between the two rubrics, but certain areas were only addressed by one or the other. Examples are provided in Table 4.
Table 4 - CLASS and Framework for Teaching Behaviors and Corresponding Student Survey Questions

<table>
<thead>
<tr>
<th>CLASS</th>
<th>Framework for Teaching</th>
<th>Student Survey Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>The teacher conveys interest in the content and materials to the students through his/her facial expressions, tone, and other non-verbal communication.</td>
<td>High levels of student engagement and teacher passion for the subject create a culture for learning.</td>
<td>My teacher is enthusiastic about the subject.</td>
</tr>
</tbody>
</table>
| Rules and behavior expectations are clearly stated or understood by everyone in the class. | Standards of conduct are clear. | My teacher explains how we are supposed to behave in class.  
I understand the rules for behavior in this class. |
| The teacher can answer all levels of student questions. | N/A | My teacher is able to answer students’ questions about the subject. |
| N/A | Students contribute to the seamless operation of classroom routines and procedures, for transitions, handling of supplies, and non-instructional duties. | Students help the teacher with classroom tasks (passing out papers, materials, etc.) |

Overall, these procedures led to the development of 64 survey questions that all have a basis in either overlapping areas of literature reviews or are grounded in descriptions of teacher behaviors from valid observational rubrics. The resulting set of questions represents both behaviors that were found in all sources as well as behaviors that may have been unique to a particular evaluation instrument (see Table 4). Since the constructs come from several different sources, there is no single grouping structure to draw upon when organizing questions. Due to the fact that these new constructs have not previously been tested for their reliability, the investigation will include both an internal reliability index using the constructs listed as well as an exploratory factor analysis to determine optimal alignment.

Though shown to be valid measures of teacher quality, the nature of the questions within a student feedback survey can affect its predictive ability. Murray (1983) investigated the questions that asked about both high-inference and low-inference teacher behaviors on student surveys. High-inference questions such as “Is the instructor clear?” or “Does the teacher plan a good lesson?” are not able to communicate information about actual teacher behaviors in a classroom. On the contrary, questions regarding low-inference behaviors require less judgment on the part of the observer, thus allowing students to rate them more objectively. Instead of asking about instructor clarity, a related question concerning a low-inference behavior might ask the student to rate the frequency of whether “My teacher uses examples or illustrations to help
explain ideas”. By asking questions about concrete behaviors that are easy to identify in addition to asking about the frequency of behavior, the validity and reliability of student surveys improves. The survey therefore uses a rating scale from 1 to 5 that asks about the frequency of teacher behaviors. The rating scale categories include ‘Every Time’, ‘Almost Always’, ‘Often’, ‘Sometimes’, and ‘Never’.

Further content validity was provided through interviews with both former teachers and content experts. First, five former teachers were interviewed and asked to read the question, describe whether the question was understood, state what objective the question is likely trying to address, and finally, provide an assessment of how well the question addressed that objective. Following these interviews, several questions were revised, elaborated, or eliminated based on clarity and ability to match survey question with intended objective. Additionally, four content experts were provided with the survey and asked to provide feedback on whether the questions covered an adequate range of teacher behaviors, whether the questions were asking about important teacher behaviors, and how questions might be improved. Again, questions were revised based on this feedback.

Cognitive Interviews

The next step in the validation process includes three rounds of cognitive interviews. Cognitive interviews are helpful in addressing common threats to validity associated with surveys (Porter et al., 2010; Desimone & Le Floch, 2004). Threats to survey validity arise due to complex phenomena being asked about, respondents answering in socially desirable ways, or respondents not being clear about what questions are asking; with cognitive interviews guarding against these threats. In order to respond accurately, respondents must be able to “comprehend an item, retrieve relevant information, make a judgment based upon the recall of knowledge, and map the answer onto the reporting system” (Desimone & Le Floch, 2004, p. 6). Cognitive interviews allow the researcher to determine which part of the process respondents may be having difficulty with and why.

There are two main types of cognitive interviewing (Beatty & Willis, 2007). The first, referred to as a ‘think-aloud’, allows respondents to verbalize the mental process as they read and answer each question. The second style takes a more active approach on the part of the researcher in which respondents are asked specific questions about survey items. The current investigation will draw upon both interview types as they each offer certain advantages. Respondents will use the think-aloud style as they first encounter each question in order to assess overall question clarity. There will also be specific instructions to describe what teacher behaviors or experiences they are drawing upon when answering the question. If students are drawing on unintended teacher behaviors, follow-up questions will be asked about why the student chose these behaviors. There will also be specific questions about items that are identified by the researcher as potentially confusing or ask about constructs that were challenging to translate into survey questions.
In the first round, 10 students were interviewing at a local private high school in Nashville, TN. Instructions and questions that are confusing or questions that are interpreted in ways that do not address the teaching objective will be revised and tested again in a second round of cognitive interviews with 15 students at schools in Georgia. These first two rounds represent a more exploratory analysis that focuses on exposing a full range of possible problems (Blair & Brick, 2009). Following further revision, a final round of cognitive interviews intended to be more confirmatory will be conducted with 15 additional students in Georgia.

Pilot Testing

Sample

There will be two rounds of pilot testing to assess the internal consistency and predictive validity of the survey. The first round of pilot testing is taking place at a local high school in Nashville, TN, with surveys distributed to 140 students. Internal consistency within construct and sub-domain will be assessed, with additional tests run to check for unusually large variance within teachers and possible differences in answer pattern.

The second round of pilot testing is planned for Spring 2011. The bulk of test validation work will be conducted in Georgia as part of the state’s Race to the Top initiative. The sampling frame includes seven districts that represent urban, suburban, and rural districts, with basic information on each district provided below. Schools will be selected based on

| Table 5 – 2008-2009 Demographic Information for Proposed Districts in Georgia |
|---------------------------------|-----------------|--------------------------|--------------------------|-----------------------------|
| Dekalb                          | 99,775          | 65,936                   | 7,436                    | Suburb: Large               |
| Griffin-Spalding                | 10,823          | 7,221                    | 105                      | Suburb: Large               |
| Hall                            | 25,629          | 13,714                   | 4,477                    | Rural: Fringe               |
| Meriwether                      | 3,449           | 2,794                    | 30                       | Rural: Distant              |
| Pulaski                         | 1,593           | 957                      | 21                       | Town: Distant               |
| Rabun                           | 2,244           | 1,362                    | 131                      | Rural: Remote               |
| Savannah-Chatham                | 33,994          | 21,024                   | 617                      | City: Mid-size              |

The final site pilot location is Westside Consolidated School District in Craighead County, Arkansas. Westside is rural, primarily white, and small, consisting on only one middle and high school respectively. Descriptive statistics for the district are displayed in Table 6 below.
Table 6 - Descriptive Statistics for Westside Consolidated School District

<table>
<thead>
<tr>
<th></th>
<th>Middle School</th>
<th>High School</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Students</td>
<td>351</td>
<td>633</td>
</tr>
<tr>
<td>Number of White Students</td>
<td>343</td>
<td>617</td>
</tr>
<tr>
<td>Number eligible for Free Lunch</td>
<td>146</td>
<td>221</td>
</tr>
<tr>
<td>Number eligible for Reduced Lunch</td>
<td>37</td>
<td>63</td>
</tr>
</tbody>
</table>

**Measures**

*Student Achievement*

A teacher’s value-added scores for each classroom will be calculated using a variety of approaches. These are discussed in more detail below.

*Academic Engagement*

Student engagement examines student’s report on their interest in learning. The measures for the current investigation were developed and tested by the Consortium on Chicago School Research. The 4-point Likert scale ranges from ‘Strongly Agree’ to ‘Strongly Disagree’ and includes six questions. Overall summary statistics for high school include individual separation (1.37), individual level reliability (.65) and school level reliability (.88). Item characteristics of are provided below.

Table 7 - Measure of Academic Engagement

<table>
<thead>
<tr>
<th>Item</th>
<th>Item Difficulty</th>
<th>Item Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>The topics we are studying are interesting and challenging</td>
<td>0.54</td>
<td>0.71</td>
</tr>
<tr>
<td>I am usually bored in this class</td>
<td>0.76</td>
<td>0.89</td>
</tr>
<tr>
<td>I usually look forward to coming to this class</td>
<td>0.76</td>
<td>0.57</td>
</tr>
<tr>
<td>I work hard to do my best in this class</td>
<td>-0.37</td>
<td>0.88</td>
</tr>
<tr>
<td>Sometimes I get so interested in my work I don’t want to stop</td>
<td>0.93</td>
<td>0.75</td>
</tr>
<tr>
<td>I often count the minutes until class ends</td>
<td>1.18</td>
<td>1.07</td>
</tr>
</tbody>
</table>

*Academic Efficacy*

Academic efficacy refers to student perceptions of their competence to do their class work. It was developed as part of the Patterns for Adaptive Learning Scales (PALS) survey at the University of Michigan. The five question scale uses the same 4-point Likert ratings as academic engagement, and has a Cronbach alpha score of .78.

---

1 [www.nces.ed.gov](http://www.nces.ed.gov)
Analytical Strategy

Relationship to Teacher Value-Added

The relationship to a teacher’s value-added scores will help provide evidence for criterion validity as gains in student achievement are arguably the most common metric for performance in education. To calculate value-added scores for teachers, a model adapted from a similar study relating scores on an observational rubric with a teacher’s value-added will be employed (Kane, Taylor, Tyler, & Wooten, 2010). Model 1.1 includes the achievement of student $i$ of teacher $k$ as the outcome, with a teacher’s total score within a classroom, a student’s prior achievement, student characteristics that may influence achievement (free and reduced price lunch status and race), and teacher experience. The error terms represent unexplained variance at the student ($\varepsilon$) and classroom level ($v$). It is possible that classroom characteristics may influence student ratings, so classroom variables ($X_{jk}$) will be included.

$$A_{ijk} = \beta_0 + \beta_1 TS_{jk} + \beta_2 X_{ijk} + \beta_3 Exp_k + \beta_4 X_{jk} + \varepsilon_{ijk} + v_{jk}$$

In the addition the model described above, results will investigate whether the relationship between total student score and a teacher’s value-added is dependent on value-added model choice. Common models that will be employed include a general persistence model, a variable persistence model, a one-level fixed effects approach, and a residual approach (McCaffrey, Han, & Lockwood, 2009).

Variance Component Analysis

An analysis of variance investigates student responses of the same teachers and of different teachers to create a discrimination index. Assuming there are differences among frequency of teacher behaviors, ratings of the same teacher should be more similar than ratings of different teachers. Subsequently, the within group variance would ideally be small relative to the between group variance. Therefore, the within and between variance of ratings for both total scores and for each scale will be calculated and compared using an F-test.

Factor Analysis

While there is a strong connection between survey constructs and previously validated scales, the fact that the constructs were developed by the researcher suggests that both a confirmatory and exploratory factor analysis is appropriate. First, the existing scales will be tested for internal consistency using Cronbach alphas. Cronbach alpha measures how closely a set of items are related together as a group. Generally, alpha levels above .7 indicate that items have adequate internal consistency. Further, an exploratory factor analysis will be completed that identifies constructs suggested by the data.
Item Reliability

Item discrimination provides additional evidence of survey reliability by measuring the relationship between individual items and a teacher’s total score. Items that have either no relationship or a negative relationship may undermine validity as the item may be measuring something other than intended. Item discrimination will be calculated using a Spearman correlation between item score and a teacher’s total score. This test is preferable to using Pearson’s correlation because it is unknown whether the relationship between each question and the total score should be expected to be linear.

Possible Threats to Validity

There are potential factors that may detract from the survey validation. First, it is possible that students may not spend adequate time answering survey questions. This could result in students putting random answers that may have no relationship to the actual frequency of teacher behavior. To prevent this, answers that fall 1.5 standard deviations away from the class mean will be flagged. Though this discrepancy may have meaning at the individual question level (for instance, if a teacher did not check for understanding with all students), a repeated pattern of deviance from the class mean may indicate that the student was not taking the survey seriously. Therefore, students who have more than 1/3 of their answers flagged will be checked for repeated, consecutive answers or suspicious answer patterns.

Next, a possible threat to validity is the time that a child spends in a teacher’s classroom. A student may have a biased opinion of a teacher if they have not had adequate time to observe the variety of behaviors that are asked about in the survey. While there is no specified minimum number of days that a student needs to attend to observe a full range of a teacher’s behaviors, it is reasonable to assume that a student has had enough time to observe the teacher if they have spent more than a month in their classroom as the behaviors listed on the survey should be observed on a regular basis. The survey will therefore include a preliminary question that asks the students how long they have been enrolled in this teacher’s class. Students that answer ‘less than 1 month’ will be excluded when calculating a teacher’s total score.

A further threat would be that characteristics of both the students and the teachers may influence ratings. For instance, there is some evidence that students rate female teachers higher (Aleamoni, 1999). Alternatively, there is also evidence of an effect on student achievement when teachers and students have the same race (Dee, 2004). To prevent both of these circumstances, controls for teacher gender, race, ethnicity, experience, and advanced degrees will be included. Additionally, student level controls for gender, race, ethnicity, and socioeconomic status will be part of the model. As a control for race matching, a dummy variable will be included that indicates whether the student race and teacher race are the same.

Additionally, a variable that may bias results is class size. If teachers behave differently because of class size, or if students rate teachers differently when in larger classes, the results
could be biased. Class size will be included as a control variable as determined by the number of students that are assigned to a teacher’s classroom. This represents a more preferred circumstance than number of students answering the survey because this is dependent on attendance for the day the survey is given.

**Results**

**Relationship to Outcome Measures**

The three outcomes used in this investigation include two measures that were administered concurrently with the survey as well as a teacher’s value-added scores. At the time of this writing, achievement scores have not yet been provided by the state of Georgia so value-added has not been calculated. The two concurrent measures are a 6-question index of academic engagement and a 5-question index of academic self-efficacy. For the first outcome, correlations between a teacher’s total score, each of the scale scores and academic engagement as well as academic self-efficacy are displayed in Table 8 below.

<table>
<thead>
<tr>
<th>Teacher Total Score</th>
<th>Presenter</th>
<th>Manager</th>
<th>Counselor</th>
<th>Coach</th>
<th>Motivator</th>
<th>Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic Engagement</td>
<td>0.781</td>
<td>0.767</td>
<td>0.570</td>
<td>0.738</td>
<td>0.740</td>
<td>0.832</td>
</tr>
<tr>
<td>Academic Self-Efficacy</td>
<td>0.651</td>
<td>0.655</td>
<td>0.511</td>
<td>0.574</td>
<td>0.622</td>
<td>0.640</td>
</tr>
</tbody>
</table>

Both total scores and each of the scale scores for teachers show a very strong relationship to each of the two outcomes with all correlations being positive and significant. For engagement, the data follows an intuitive pattern, with the strongest relationship coming from a teacher’s ability to motivate students and the weakest relationship between classroom management and engagement. The correlations are slightly higher for engagement than academic self-efficacy.

**Survey Properties**

**Internal Consistency**

The existing scales will be tested for internal consistency using Cronbach alphas. Cronbach alpha measures how closely a set of items are related together as a group. Generally, alpha levels above .7 indicate that items have adequate internal consistency. Table 9 displays the
Cronbach alpha scores for each scale as well as the number of items. Overall, 5 of the 6 scales display the desired levels of internal consistency, suggesting that questions within each construct are measuring similar aspects of teacher quality.

Table 9 - Cronbach Alpha Values for Survey Scales

<table>
<thead>
<tr>
<th></th>
<th>Presenter</th>
<th>Manager</th>
<th>Counselor</th>
<th>Coach</th>
<th>Motivator</th>
<th>Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cronbach Alpha Score</td>
<td>0.893</td>
<td>0.560</td>
<td>0.821</td>
<td>0.824</td>
<td>0.850</td>
<td>0.820</td>
</tr>
<tr>
<td>Number of Items</td>
<td>13</td>
<td>9</td>
<td>10</td>
<td>13</td>
<td>10</td>
<td>8</td>
</tr>
</tbody>
</table>

Variance Component Analysis

Assuming there are differences among frequency of teacher behaviors, ratings of the same teacher should be more similar than ratings of different teachers. Subsequently, the within group variance would ideally be small relative to the between group variance. All questions had F-values that were significant, indicating that all questions had smaller within teacher variance.
Works Cited


Rockoff, J., Jacob, B., Kane, T., & Staiger, D. (2008). *Can You Recognize an Effective Teacher When You Recruit One?* NBER.


