Quality Assurance in Education

Designing non-cognitive construct measures that improve mathematics achievement in Grade 5-6 learners: A user-centered approach

Madhabi Chatterji, Meiko Lin,

Article information:

To cite this document:
Madhabi Chatterji, Meiko Lin, (2018) "Designing non-cognitive construct measures that improve mathematics achievement in Grade 5-6 learners: A user-centered approach", Quality Assurance in Education, Vol. 26 Issue: 1, pp.70-100, https://doi.org/10.1108/QAE-11-2017-0081
Permanent link to this document:
https://doi.org/10.1108/QAE-11-2017-0081

Downloaded on: 26 April 2018, At: 09:16 (PT)
References: this document contains references to 81 other documents.
To copy this document: permissions@emeraldinsight.com
The fulltext of this document has been downloaded 33 times since 2018*

Received 15 November 2017
Revised 20 November 2017
Accepted 10 December 2017

Designing non-cognitive construct measures that improve mathematics achievement in Grade 5-6 learners

A user-centered approach

Madhabi Chatterji and Meiko Lin
Teachers College, Columbia University, New York, New York, USA

Abstract

Purpose – The purpose of this study was to design and iteratively improve the quality of survey-based measures of three non-cognitive constructs for Grade 5-6 students, keeping in mind information needs of users in education reform contexts. The constructs are: Mathematics-related Self-Efficacy, Self-Concept, and Anxiety (M-SE, M-SC, and M-ANX).

Design/methodology/approach – The authors applied a multi-stage, iterative and user-centered approach to design and validate the measures, using several psychometric techniques and three data samples. They evaluated the utility of student-level scores and aggregated, classroom-level means.

Findings – At both student and classroom levels, replicated evidence supported theoretically-grounded validity arguments on information produced by four of five scales tapping M-SC, M-ANX and M-SE. The evidence confirmed a second order, two-factor structure for M-SC, representing *positive math affect* and *perceived competence*, and a one factor structure for M-ANX representing *negative math affect*. Consistent with the literature, these served as precursors to a *perceived confidence* factor of M-SE which, in turn, positively influenced mathematics achievement scores, off-setting negative effects of M-ANX. Research is continuing on a *self-regulatory efficacy* factor of M-SE, which yielded mixed results.

Practical implications – The survey scales are in line with current reform policies in the United States calling for schools to monitor changes in cognitive and non-cognitive domains of student development. Validated scales could be useful in serving information needs of teachers, decision-makers and researchers in similar school-based contexts.

Originality/value – This study demonstrates a comprehensive, user-centered methodology for designing and validating construct measures, departing from purely psychometric traditions of scale development.

Keywords Validity, Structural equation modeling, User-centered design, Math achievement, Math self-efficacy, Education reforms

Paper type Research paper

Purpose

There is mounting research evidence suggesting that cognitive and non-cognitive capacities of learners develop reciprocally in given domains and are inter-connected (Seaton *et al.*, 2014). The purpose of this study was to design and formally validate self-reported measures

The authors acknowledge the thoughtful comments by Michael T. Kane of Educational Testing Service, and feedback on earlier versions of the paper from reviewers and discussants at annual conferences of the National Council on Measurement Education.

Quality Assurance in Education
Vol. 26 No. 1, 2018
pp. 70-100
Emerald Publishing Limited
0968-4883
DOI 10.1108/QAE-11-2017-0081

of selected, non-cognitive constructs for fostering student learning in mathematics domains targeted by current education reform initiatives in the United States. We use the term "non-cognitive" here to refer to dispositions, beliefs, social-emotional or affective mind sets, and attitudinal constructs (Duckor, 2017).

The study's specific focus was on measuring Mathematics-related Self-Efficacy (M-SE), Self-Concept (M-SC) and Anxiety (M-ANX) at the student level, which could then be aggregated at classroom and school levels to monitor progress over time, as emphasized in reform initiatives. To that end, the study employed a user-centered design and validation approach guided by, and situated in, the projected contexts of assessment information use. Specifically, the construct measures were validated at:

- the individual student level, for serving information needs of teachers and students in classroom learning contexts; and
- the classroom levels, for supporting research-related or evaluative decision-making needs at upper levels of the education system.

Theoretical and policy rationale

Current education reform movements in the United States such as the Every Student Succeeds Act (ESSA) and the immediately preceding Common Core State Standards Initiative, provided a policy rationale for this work (Porter *et al.*, 2011; Ravitch, 2011; see: https://ed.gov/policy/elsec/leg/essa/index.html). These reforms, continuing since the passage of the No Child Left Behind Act of 2002, call for higher standards and improved student outcomes in mathematics from the earliest years of schooling, alongside student development in non-cognitive domains.

Duckor (2017, pp. 61-62) recently commented on the dearth of appropriately designed and validated non-cognitive instruments suitable for school-based applications. Such observations, coupled with the ESSA's rhetoric on monitoring school improvement on "non-cognitive" indicators other than student achievement, point to an immediate need for validated measures that can foster student achievement in classrooms, while supporting school improvement and accountability initiatives in school and district systems.

Simultaneously, there is also a need to deepen scientific understandings of the potential influences of non-cognitive constructs like M-SE, M-SC and M-ANX on students' mathematics achievement through contextually based studies using instruments designed for school settings. A growing body of psychological and cognitive research points to the reciprocal manner in which individuals develop in cognitive and non-cognitive domains in given competency areas (Seaton *et al.*, 2014; Usher and Pajares, 2008). Such evidence suggests that as students experience greater success with particular mathematics tasks, their M-SE and M-SC levels would likely improve, with M-ANX levels reducing. Reciprocally, as these non-cognitive attributes develop in the desired directions, students' mastery experiences in mathematics domains will also improve. With added research from applied settings, evidence-based guidelines for supporting affirmative student development in classrooms, can follow.

The validation arguments guiding the present research were grounded in the literature on "Sources of Self-Efficacy" (Joët *et al.*, 2011; Usher and Pajares, 2008, 2009). This body of literature suggests that individuals are not born with innate levels of attributes like M-SE, M-SC and M-ANX. Rather, these are largely learned mind-sets, attitudes and beliefs that are acquired based on one's accumulated life experiences. The "Sources of Self-Efficacy" literature speaks to the purported precursors of self-efficacy beliefs that individuals could also develop, which in turn could influence their self-efficacy levels and future successes in particular domains, including mathematics.

Usher and Pajares (2008, p. 755) noted in their review of the literature on the Sources that "not all researchers have been attentive to issues related to construct validity or to theoretical guidelines related to the nature of the sources" in their studies. We approached the present study with attention to these concerns.

Other educational researchers have also called for the adoption of a formal "construct validation" approach for better investigation of properties of non-cognitive construct measures that rely on self-reported surveys (Pajares and Miller, 1994, pp. 363–365). Biases stemming from item content similarity and specificity during construct measurement, or the frames of reference that individuals use when they respond to surveys are enduring areas of concern. We attempted to address these particular issues through our procedures.

Lastly, we acknowledge that there are long-standing research interests in the constructs we attempted to measure and investigate here (see for examples: Ahmed *et al.*, 2012; Galla and Wood, 2012; Gutman and Schoon, 2013; Hoffman, 2010; Lee, 2009; Marsh and Martin, 2011; Marsh *et al.*, 1997; Mullen and Schunk, 2011; Pajares and Graham, 1999; Pajares and Miller, 1994; Perry *et al.*, 2007; Schunk and Usher, 2011; Vick and Packard, 2008; Wang *et al.*, 2012). However, a majority of these prior studies and existing instruments focused on students at the middle and high school levels, or on teachers, college-going individuals and adult learners. This left a small area open for further research with preadolescent, elementary school students that we decided to undertake.

User-centered measurement rationale

Once designed, most instruments – whether they are student achievement tests or survey-based measures tapping non-cognitive attributes – are used for serving more than one inferential need and purpose in applied settings. The most recent publication of the *Standards for Educational and Psychological Testing*, therefore, underscores that all intended score-based inferences from tests and measuring tools must be validated in the anticipated contexts of use (American Educational Research Association [AERA], American Psychological Association [APA] and National Council on Measurement in Education [NCME], 2014). Particularly, Standard 1.2 on Validity states that "A rationale should be presented for each intended interpretation of test scores for a given use", with appropriate evidence and theoretical support for the use of construct measures in each application context (p. 23, see also, Cronbach, 1971; Kane, 2006, 2013; Messick, 1989).

Methodological frameworks that tie together instrument design and validation processes with the assessment users' needs for information, are still under-developed or limited in the measurement literature (Mislevy, 2006 in Brennan, 2006). To address this gap, the present study employed a user-centered process for achieving the ends underscored by Standard 1.2.

The Process Model, demonstrated in Figure 1 as applied in this study, is a user-centered and integrative methodology (Chatterji, 2003). Assessment design begins with a specification of the assessment user context, along with the intended inferences and uses to be made with assessment results. The instrument (including all sub-scales and construct measures) is then designed in Phases II-III guided by the context specifications. Validation then involves cycles of evidence gathering and evaluation to determine the extent to which the scales/instruments designed produce the caliber of data necessary to support each intended interpretation and use. Improvements are made to the items, scales and measures using evidence from each cycle as shown in Phases IVA-B in Figure 1. Relevant kinds of evidence – content-based, empirical, qualitative and quantitative – are examined. To optimize the utility of measures in the desired contexts, evidence-gathering and instrument refinements continue until necessary evidence of validity, reliability and overall psychometric quality is obtained for the information needs and uses specified (Chatterji, 2003, *in press*).

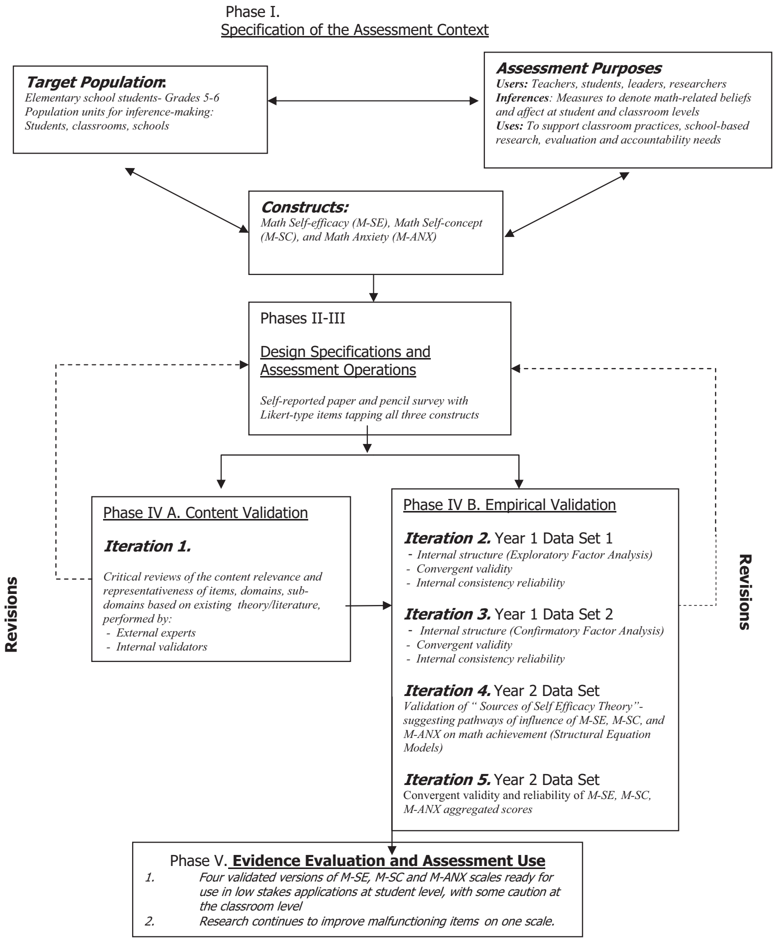

Phase I.
Specification of the Assessment Context
Target Population:
Elementary school students- Grades 5-6
Population units for inference-making:
Students, classrooms, schools
Assessment Purposes
Users: Teachers, students, leaders, researchers
Inferences: Measures to denote math-related beliefs and affect at student and classroom levels
Uses: To support classroom practices, school-based research, evaluation and accountability needs
Constructs:
Math Self-efficacy (M-SE), Math Self-concept (M-SC), and Math Anxiety (M-ANX)
Phases II-III
Design Specifications and Assessment Operations
Self-reported paper and pencil survey with Likert-type items tapping all three constructs
Revisions
Phase IV A. Content Validation
Iteration 1.
Critical reviews of the content relevance and representativeness of items, domains, sub-domains based on existing theory/literature, performed by:
- External experts
- Internal validators
Phase IV B. Empirical Validation
Iteration 2. Year 1 Data Set 1
- Internal structure (Exploratory Factor Analysis)
- Convergent validity
- Internal consistency reliability
Iteration 3. Year 1 Data Set 2
- Internal structure (Confirmatory Factor Analysis)
- Convergent validity
- Internal consistency reliability
Iteration 4. Year 2 Data Set
Validation of " Sources of Self Efficacy Theory"- suggesting pathways of influence of M-SE, M-SC, and M-ANX on math achievement (Structural Equation Models)
Iteration 5. Year 2 Data Set
Convergent validity and reliability of M-SE, M-SC, M-ANX aggregated scores
Revisions
Phase V. Evidence Evaluation and Assessment Use
1. Four validated versions of M-SE, M-SC and M-ANX scales ready for use in low stakes applications at student level, with some caution at the classroom level
2. Research continues to improve malfunctioning items on one scale.

Figure 1. User-centered assessment design and validation methodology applied to M-SE, M-SC and M-ANX measures

Theoretical frameworks for operationalizing constructs

We now discuss the literature that informed the operationalization of the three constructs in the present study, M-SE, M-SC and M-ANX, indicating how our conceptualizations differ from existing instruments and relate to the literature on the Sources of Self Efficacy (Bandura, 1997, 2006; Usher and Pajares, 2008). This section is referenced again during discussion of results.

Mathematics-related Self-Efficacy and Self-Regulatory Efficacy

Bandura (1986, 1997, 2000, 2006) defined self-efficacy, or perceived self-efficacy, as a belief in one's ability to perform tasks required to accomplish goals and gain control over one's life's circumstances. Closely related to a sense of self-agency, self-efficacy beliefs are based on Bandura's (1997) social cognitive theory and have been shown to influence our actions, effort, perseverance, resilience to adversity, self-regulation and independence leading to the realization of goals, in a number of areas. Individuals develop "self-efficacy beliefs that become instrumental to the goals they pursue and to the control they are able to exercise over their environment" (Pajares, 2002, p. 116).

Empirically, self-efficacy is known to be a positive predictor of individual performance-related behaviors and explains one's capabilities to accomplish specific tasks and goals in a domain. In academic domains like mathematics, self-efficacy is a student's belief that he or she can succeed in completing academic tasks in specific courses of study (Bong and Skaalvik, 2003; Ferla *et al.*, 2009).

Closely connected to self-efficacy beliefs are self-regulatory practices (Bandura, 1986, 1997; Pajares, 2002). Previous studies show that students' beliefs in regulating their own learning and mastery of different academic subjects affect their levels of motivation and their experienced academic success (Bandura, 1993; Zimmerman, 1989, 1994; Zimmerman and Bandura, 1994; Zimmerman and Martinez-Pons, 1990). Further, students with high self-efficacy are more capable of using more effective self-regulatory strategies to monitor their academic work and maintain high academic achievement (Pajares, 2002).

According to Bandura (1986), perceived *Self-Regulatory Efficacy* may be more relevant than perceived self-efficacy to achieve desired results in familiar and regularly performed activities in some domains. In a detailed guide to developing self-efficacy scales in health-related domains, Bandura (2006, p. 311; parenthesis added) states:

> Many areas of functioning are primarily concerned with self-regulatory efficacy to guide and motivate oneself to get things done that one (already) knows how to do. In such instances, self-regulation is the capability of interest. The issue is not whether one can do the activities occasionally, but whether one has the efficacy to get oneself to do them regularly in the face of different types of dissuading conditions. For example, in the measurement of perceived self-efficacy to stick to a health-promoting exercise routine, individuals judge how well they can get themselves to exercise *regularly* under various impediments [. . .].

Bandura's (2006) above quote suggests that *Self-Regulatory Efficacy* could be viewed as a part of self-efficacy governing learning in scholastic achievement domains, where regular practice and effort are necessary. In this study, we treated *Self-Regulatory Efficacy* in mathematics as a dimension of M-SE rather than as a separate construct, guided by the reasoning that to master mathematics, students must be able to effectively self-regulate their learning processes, levels of effort, engagement and commitment to practicing mathematics.

Empirically, academic self-efficacy has been found to be related to students' future expectancies for successfully performing specific academic tasks (Wigfield and Eccles, 2000). Individuals with higher self-efficacy beliefs in a domain tend to have higher outcome expectations (Gainor and Lent, 1998; Usher and Pajares, 2008).

Individuals typically undertake tasks in which they believe they will be capable and avoid those in which they believe they will not succeed (Bandura, 1997; Vrugt, 2004). Studies on M-SE have found that the construct is a strong mediator of *choices* individuals make regarding their involvement in academic work, ultimately influencing their achievement levels (Bandura, 2000; Roeser *et al.*, 2000; Usher and Pajares, 2009). This literature suggests that students with high M-SE would likely choose to engage more deeply in mathematics tasks, and therefore, they would also be likely to show high achievement levels as a consequence of that engagement.

Operationalizing self-efficacy. Self-efficacy is a domain-specific construct (Bandura, 1997, 2006). Domain specificity, as opposed to domain generality, refers to the unique properties of an attribute which are manifested only when individuals engage in tasks in a given domain but not in others. Although earlier studies had ignored Bandura's advice on domain specificity in measuring self-efficacy (Pajares and Miller, 1994), most researchers now treat self-efficacy as a domain-specific construct, as we did.

Self-efficacy beliefs are typically tapped through self-reported surveys or interviews. The phrase "I can" in items denotes self-efficacy. The construct is purported to have three measurable properties across activities and contexts (Bandura, 1986, 1997; Zimmerman, 2000): level, strength and generality. The *level* of self-efficacy refers to the degrees of self-perceived difficulty in performing specific tasks or meeting goals. The *strength* of self-efficacy reflects the certainty of a person's belief in successfully performing specific tasks. The *generality* of self-efficacy refers to the degree to which self-efficacy beliefs are transferable across other similar construct domains. The present study focused on measuring the strength of M-SE beliefs.

Based on Bandura's (2006) guide for scale construction in measuring strength of "driving self-efficacy", for example, a series of items ask how confident or certain persons are that they can drive on freeways, city roads or narrow mountain roads (Bandura, 2006). For tapping efficacy in general mathematics domains applicable to elementary school, we used parallel item structures asking how sure (or certain) students were that they could complete their mathematics homework, classwork or examinations in classroom settings.

Efficacy response scales should be unipolar, ranging from 0 to a maximum strength, and without the use of negative numbers. While Bandura (2006, p. 308) suggests a 100-point scale, with 10-point intervals from 0 ("Cannot do") to complete assurance at 100 ("Highly certain can do"), he also endorses "simpler response formats" which retain the same scale structure and descriptors but use single unit intervals. Such adaptations accommodate "variations in format depending on the age of the respondents and the sphere of efficacy" (Bandura, 2006, pp. 311–312, 323). We followed the latter recommendation, given the young age of our targeted population.

Domain specifications and indicators of M-SE for pre-adolescent students. In sum, we operationalized M-SE with two closely related dimensions as relevant to the pre-adolescent population for the present study. These reflected:

- a *Perceived Self-confidence* dimension, reflecting a "can do" spirit in accomplishing tasks in mathematics; and
- a *Self-Regulatory Efficacy* dimension for achieving mathematics learning goals, including one's persistence, effort, independence, engagement in help-seeking behaviors, leading to successful goal achievement.

Five specific indicators fell under these sub-domains, as follows:

Perceived Confidence in mathematics.

- Students believe that they "can do" math required at grade level (self-beliefs in success in domain).
- Students report confidence in succeeding in math tasks in future (positive outcome expectancy).
- Students believe that they are capable of success in difficult math tasks (self- beliefs in success despite challenges or against odds).

Self-Regulatory Efficacy in learning mathematics.

- Students report using self-regulatory habits to accomplish math tasks (persistence, resilience, help- and solution-seeking behaviors).
- Students report independence in giving self-motivated effort in completing math tasks (effort, independence).

An empirical question remained as to whether the item composition under these two dimensions would hold up as consistently separate, accounting for distinctly unique variance components. Existing scales on which the two are treated separately are labeled as "academic self-efficacy" versus "self-regulatory efficacy", respectively (Bandura, 2006; Usher and Pajares, 2008).

Survey items were written tied to the indicators to generate a 21-item pool. Based on pilot testing of the preliminary instrument with Grade 4-6 students in one school, we arrived at a three-point, unipolar response scale for all items with Yes (three points, very sure I can do this), Not sure (two points, not sure I can do this) and No (one point, I cannot do this) responses. Compared with five- or seven-point scales we also tested, children responded most meaningfully and consistently to the three-point range of options.

Although our response scale structure differs significantly from the 100-point scale in Bandura's (2006) guide, we were able to test its viability across multiple samples in the current study. All items were written as first-person statements or questions to facilitate readability and meaningfulness for the target population.

By design, we chose to write M-SE items tied to general mathematics, consistent with the way the subject is typically taught at the elementary school levels. This gave the survey higher utility across various mathematics instructional units taught in Grades 5-6, such as, long division, geometry or prime numbers. Appendix 1 shows all M-SE items (Table AI).

Sources of Self-Efficacy theory

Usher and Pajares (2008) performed a review of the literature on the Sources of Self Efficacy beliefs based on an analysis of Bandura's (1986, 1997, 2006) social cognitive theory and past empirical literature. They identified four major Sources of Self efficacy beliefs, characterizing these as *antecedent* factors reflecting one's interpretations of prior life experiences that purportedly lead to the development of self-efficacy beliefs. These notions partly informed our item development procedures and relational arguments to validate pathways of influence. The four Sources are:

(1) *Mastery experience*: Individuals interpret *mastery* based on their levels of past success or failure at given tasks vis-à-vis the amount of effort and time they believe they expended. Applied to areas like mathematics mastery at school, self-efficacy beliefs develop as students interpret information from their own previous achievements and experiences of mastery in performing mathematics tasks and problems (Joët *et al.*, 2011; Usher and Pajares, 2006, 2008).

(2) *Vicarious experience*: Individuals build their efficacy beliefs through the *vicarious experience* of observing how their role models or peers perform and succeed at tasks like mathematics problems. In school, students develop beliefs of their own mathematics capabilities in relation to how they perceive others' performances. For example, a student could interpret a test score as low or high depending on how others in class performed on the same test. When most of the class perform poorly, a student's self-efficacy tends to be raised by comparison. Normative and peer-to-peer comparisons, such as getting a better grade than others, or performing as well as or better than the best students in class help affirm one's self-efficacy beliefs in an area (Usher and Pajares, 2008).

(3) *Social or verbal persuasions*: Verbal and social persuasions are words of encouragement or statements of belief regarding a person's abilities, such as those made by parents, teachers and others that a person values. In school, such "persuasions" can bolster one's confidence in academic domains. Young learners

are often not skilled at making accurate self-appraisals, and they depend on evaluative feedback and judgments about their academic abilities from others that they view as important (Usher and Pajares, 2006, 2008).

(4) *Emotional and physiological states*: Finally, according to Bandura (1986, 1997), self-efficacy beliefs are informed by one's emotional and physiological states associated with engaging in tasks in a domain. Individuals associate and interpret their physiological states, such as anxiety, stress, fatigue, excitement and mood, as indicators of competence. Strong negative emotions in response to school-related tasks like math problems serve as cues for negative outcome expectancy. While a little anxiety might help, high anxiety can undermine self-efficacy by raising self-doubts as to whether one will succeed. General feelings of well-being and reduced negative emotional states, on the other hand, tend to strengthen self-efficacy beliefs in a performance domain (Usher and Pajares, 2008).

According to Bandura (1986, 1997), the Sources of Self -efficacy (Sources, hereafter) have a causal influence on one's self-efficacy beliefs in a domain. Causality and correlational influences have been investigated and inferred by researchers with linear regression-based approaches and variations thereof, including path analysis and structural equation modeling (SEM). Experimental designs, however, have rarely been used. A few studies have included covariates as control variables, such as prior ability levels or gender when investigating the relationships between the Sources and self-efficacy beliefs in an area (Gainor and Lent, 1998). Others have used SEM to investigate the effects of background factors like gender and learning disability status on the Sources (Hampton and Mason, 2003).

Gender and prior ability levels have been shown to yield differences on M-SE beliefs. Researchers have observed that using prior ability, test scores or grades to represent mastery experience (one Source) could interfere with the assessment of the influences of the Sources on both self-efficacy measures and achievement, when these serve as dependent variables in models. Mastery experience, measured as above, tends to be highly correlated with achievement indices (Usher and Pajares, 2008, 2009).

Mathematics self-concept: a possible source of self-efficacy in math?

Self-concept is described in the literature as another non-cognitive attribute that is a predictor of choices that individuals make regarding their involvement in specific types of academic work and performance in particular domains (Pajares and Miller, 1994). *Academic self-concept* is defined as an individual's evaluation of his or her own abilities in a specific academic area.

Sometimes also referred to as self-esteem (Bandura, 2006), it refers to self-perceptions formed through past experiences within an academic environment, based on appraisals of one's academic performance as communicated by others, or through self-appraisals drawing on social and normative comparisons. Typically also measured with surveys, academic self-concept tends to be defined at a higher level of generality than self-efficacy, with the latter reflecting more goal-specific or task-referenced evaluations by the self (Pietsch *et al.*, 2003). For example, a typical mathematics self-concept item is "I am good at mathematics".

Although they are similar, in that both predict performance to different degrees and are treated as multidimensional and domain-specific constructs, several distinctions exist between academic self-concept and academic self-efficacy. Academic self-efficacy incorporates outcome expectancy beliefs, or one's future-oriented beliefs regarding success with specific goals. On the other hand, academic self-concept is more a reflection of one's feelings and affect about being good at something based on prior self-evaluations of performance or based on how others provide evaluations (Bong and Skaalvik, 2003).

Because academic self-concept involves evaluating one's competence against those of others, respondents rely heavily on information from past mastery experiences to make such an evaluative judgment to endorse statements like "I have always been good at math" (Bong and Skaalvik, 2003). We identified a strong substantive overlap between academic self-concept and at least two Sources of Self-efficacy discussed previously (Usher and Pajares, 2008), namely, *Mastery experience* and *Vicarious experience*.

In mathematics, Bandalos *et al.* (1995) found that secondary students' M-SC, and not their M-SE beliefs, mediated their attribution to failure and success with prior mathematic experiences. Also with a secondary student sample, Pietsch *et al.* (2003) found empirical evidence supporting two main dimensions comprising M-SC – perceived competency and affect or feelings toward mathematics. Interestingly, through factor analysis, they found some overlap in item loadings on the "competency" factor across M-SE and M-SC domains.

Operationalizing Mathematics-related Self-Concept. A more concrete and observable framework for measuring academic self-efficacy versus academic self-concept was put forth by Pajares and Schunk (2002) and later by Bong and Skaalvik (2003). They suggested that differently worded questions need to be asked to tap each. Academic self-efficacy items are constructed by asking "I can" questions or statements, whereas academic self-concept is tapped by asking questions about "being good at" and "feeling good about" something. For example, "I can do the problems on math tests" would be an academic self-efficacy item, while "I am good at math" or "I enjoy doing math" would be academic self-concept items.

Also, while both constructs incorporate an individual's cognitive self-appraisals of competence in given domains, most definitions of self-concept incorporate normative information based on comparisons one makes with peers or others regarding one's performance (Bong and Clark, 1999). Marsh (1999) provides an example of a typical academic self-concept item that uses social comparisons: "I am one of the best students in my class" (as cited in Ferla *et al.*, 2009).

Domain specifications and M-SC indicators. Given the above literature, we specified the domain for M-SC with two main subdomains, representing *Positive Affect* and *Perceived Competence* in mathematics. Altogether ten items comprised the M-SC domain, including items using normative comparisons. Appendix 1 shows items tied to each subdomain.

Positive Affect toward mathematics.

- Students report they like math and find math tasks enjoyable.

Perceived Competence in mathematics.

- Students believe they are good at math based on past experiences or normative, social comparisons.

Again, we used a three-point, unipolar response scale for all items with Yes (three points, I agree this is true), Not sure (two points, not sure if this is true) and No (one point, I do not agree this is true) responses based on the pilot test. As before, items were written as first-person statements or questions. See Appendix 1 for M-SC items (Table AI).

Mathematics Anxiety: Another source of M-SE?

Negative psychological states, such as, anxiety, stress, fatigue and depressed moods, tend to provide the *opposite* information about self-perceptions – as such, they are expected to be negative correlates and predictors of both M-SC and M-SE (Pajares, 1997). Academic anxiety refers to one's physio-emotional reactions when thinking about specific subjects, such as one's worries and fears about failing in mathematics (Lee, 2009). Anxiety heightens the feelings of tension that could interfere with concentration and memory. Survey-based measures of anxiety have been found to manifest as a negative correlate of academic success in a large number of empirical studies (Jain and Dowson, 2009).

Based on extensive examination of 151 studies on mathematics anxiety, Hembree (1990) concluded that mathematics anxiety is related to both poor performance on mathematics tests and negative attitudes toward mathematics learning. High mathematics anxiety, in particular, has been shown to have deleterious effects on mathematics learning outcomes (Ashcraft, 2001, 2002; Hembree, 1990).

In academic endeavors, therefore, while anxiety can impede learning and performance, academic self-efficacy and academic self-concept are expected to do just the reverse and may even play a compensatory role. Research suggests high self-efficacy creates a feeling of calmness or serenity when approaching difficult tasks while low self-efficacy may result in an individual perceiving a task as more difficult than reality, which, in turn, may create anxiety, stress and a narrower idea on how best to solve a problem or approach an activity (Bandalos *et al.*, 1995; Eccles, 2005). From a classroom practice perspective, the interception of anxiety before it develops in individuals may avoid perpetuating negative self-beliefs in mathematics, overcoming future task avoidance (Hoffman, 2010).

As with M-SC, we found commonality in the literature on anxiety and the Sources of Self-efficacy, highlighting emotional states as a precursor of self-efficacy beliefs. Drawing on that, we hypothesized that M-ANX would serve as an antecedent factor but as a negative correlate of M-SE. High M-ANX in children would lead to lower levels of M-SE, and vice versa. Likewise, we argued that M-ANX measures would have a negative relationship with M-SC. Simultaneously, the possibility that anxiety might have a curvilinear relationship with achievement also had to also be borne in mind.

Operationalizing M-ANX in pre-adolescents. For measurement purposes, M-ANX can be treated as either a uni- or multi-dimensional variable, with self-reported items reflecting negative emotional content *vis-à-vis* a subject area like mathematics.

We recognized that a reverse orientation in the language of M-ANX items yields items that appear to denote positive emotional content superficially, or content akin to that in M-SC items. For example, "I feel nervous before math class" versus "I feel calm before math class" could both serve as positively coded and reverse-coded M-ANX items. Whether these types of items remain distinct following factor analysis is an empirical question that we investigated.

Domain specifications and indicators of M-ANX. We operationalized M-ANX as expressions of negative emotions or *Negative Affect* related to mathematics learning and achievement. M-ANX was conceptualized as a single domain with both positively and negatively worded items tied to the following indicator.

Negative emotions related to math.

- Students report feelings of anxiety/fear with math tasks.

Appendix 1 provides the six items originally tied to this domain, also with a three-point, response scale with Yes (three points, I agree this is true), Not sure (two points, not sure if this is true) and No (one point, I do not agree this is true) categories. See Items 32-37 in Table AI, Appendix 1.

Methods

This section now provides details of each phase of the process model as applied in the present study. The validation studies were conducted in five iterations shown in Figure 1. Iterations 1-4 focused on obtaining preliminary and cross-validated evidence of construct validity of M-SE, M-SC and M-ANX measures in a unitary sense (Kane, 2013). This evidence was evaluated using different data sets, to gauge the quality of construct measures for making inferences at the student level. Iteration 5 focused on validating "group mean scores" for facilitating inferences at the classroom and upper organization levels of decision-making.

Phase I: Specifying the context of assessment use

Research and policy contexts for assessment information use. The survey instrument was designed in the context of a research project investigating effects of a teacher-mediated intervention on students' cognitive and non-cognitive development in mathematics (Chatterji *et al.*, 2009; Chatterji, 2012). The project provided a clear research context for researchers and participant teachers to use the information for supporting individual student development in classrooms. With legislated school reform and accountability requirements already in effect in the region, there was also a policy context for using the assessment data aggregated at both classroom and school levels for school monitoring purposes.

Specifying constructs, populations, assessment users and score-based inferences/uses. In Phase I, we began by specifying the "context of assessment use" in terms of the targeted:

- construct domains;
- assessment purposes; and
- population.

Specifying purposes required identifying the specific assessment users, inferential needs for each user, and decisions or uses intended by each user. To specify the population, we identified the population units on whom inferences would be drawn from the construct measures, and their background characteristics (Figure 1).

For this study, the primary users of the assessment data were teachers (at the individual student level), researchers and school leaders/policymakers (both at the individual student and classroom levels). Indirect users also included the school district leaders and staff, whose interest was mainly in a formative program evaluation of the intervention in participating schools.

Teachers and researchers wanted to know: Are students' M-SE, M-SC and M-ANX measures associated and affecting mathematics achievement scores of students in theoretically predicted ways? How can we use the information from achievement tests and non-cognitive construct measures to foster student development reciprocally? Administrators and leaders were inclined to ask instead: Were students' M-SE, M-SC and M-ANX levels in classrooms where the intervention was implemented, higher than in classes without the same resources? These were all "low stakes" information uses.

The proposed inferences on individual students' M-SE, M-SC and M-ANX levels were to be based on their obtained (total) scale scores *vis-à-vis* specific mathematics instructional units delivered during a term. In the years of the study, the units dealt with the topics of long division, prime numbers, geometry and pre-algebra.

At the classroom level, the proposed inferences dealt with using M-SE, M-SC and M-ANX measures as aggregated outcomes in formative "program" evaluations. Hence, our aim was to generate group mean scores using data from M-SE, M-SC and M-ANX scales that demonstrated theoretically meaningful relationships with each other, and were reliable.

Populations, characteristics and units of analysis. The specified populations for the instrument included students enrolled in Grades 5-6. The composition of samples on key demographics is shown in Appendix 2. For both years, samples were representative of the school district on gender, ethnicity and poverty for targeted grades. The medium of instruction in school is English, and most students were fluent English-speakers, readers and writers at grade level. See Appendix 2, Tables AII-AIII for demographic breakdowns of the samples.

Phases II-III. Specifying the assessment operations

Deriving indicators from existing knowledge bases. In the first step, construct domains and subdomains were operationalized in terms of general and specific indicators based on literature reviews and existing theory already discussed.

Item writing, instrument assembly with directions. Following preliminary content validation by an outside expert panel and refinement of the item pool, a multi-dimensional survey instrument was assembled with a total of 37 items. The pool was screened for a match with indicators, difficult vocabulary, double-barreled language and ambiguous content. Each subdomain represented a dimension of the targeted construct, and was intended to yield a summated total score as the "measure" (DeVellis, 2003). The final instrument structure is presented in Appendix 1:

- M-SE: 21 items organized in two broad subdomains, *Perceived Confidence* and *Self-Regulatory Efficacy in math.*
- M-SC: 10 items organized in two subdomains, *Perceived Competence* and *Positive Affect in math.*
- M-ANX: 6 items organized in one domain, with 3 incorporating reverse-oriented wording, *Negative Affect related to math.*

Data collection and frames of reference. Along with achievement tests, the survey-based M-SE, M-SC and M-ANX scales were designed to monitor students' non-cognitive development during a three-year implementation of the intervention. Participants were students and teachers in 14 classes at four elementary schools in New York, with 20 classrooms serving as comparison settings.

Students took a paper-and-pencil version of the survey in classroom settings, administered by teachers. They were asked to think about the mathematics unit they just completed to provide ratings. Directions for scoring and administration were finalized following the small-scale pilot test.

To allow for more valid domain-specific inferences from M-SE, M-SC and M-ANX measures, both external and internal frames of reference that respondents use, matter (Marsh *et al.*, 1997). In this study, the mathematics units/classwork in effect during data collection, provided an *external* frame of reference for gathering data on M-SE, M-SC, and M-ANX. Students' self-assessments of their performance on the specific unit tests served as an immediate *internal* frame of reference for reported levels of math-related affect.

The survey data were collected at the end of two school years during mathematics class, with separate student cohorts (Year 1; Year 2). The survey was administered immediately after students had completed classroom assessments in long division, geometry or pre-algebra units in 90-min blocks. Instructions and items were read aloud by the teachers, as needed, with the survey projected on a screen. Students were allowed to ask questions about any confusing survey items or instructions and were given 45 min to complete the survey.

Phase IV. Validation: samples, validation arguments, questions and analyses

Samples and data sets. Sample sizes consisted of 506 students in Year 1 and 469 students in Year 2, with complete data on all items and background variables. The breakdown of each year's sample on key demographics is shown in Appendix 2. The samples in both years were representative of the school district on gender, ethnicity and poverty for those grades.

Nomological network to guide measure-based interpretations and uses. Cronbach and Meehl (1955) referred to the conceptual network of related psychological variables, as a nomological network. A nomological network suggests a system of meanings among measured constructs based on existing science-based knowledge and understandings. Nomological networks offer theoretical frameworks that facilitate the design and validation of new or different construct measures. Once the place of a given construct in a substantive theoretical network is ascertained, it becomes plausible to formulate and empirically test hypotheses about how measures will likely behave in relation to other variables.

To examine directional pathways and relational properties of M-SE, M-SC and M-ANX measures with each other and with mathematics achievement, we drew on the Sources of Self-Efficacy theory for formulating nomological models. To be able to claim validity of measures in the projected contexts of use, we argued that (after Kane, 2006):

- the evidence should support the theorized internal scale structure, convergent validity and reliability in item sets tied to the M-SE, M-SC and M-ANX constructs at the individual student and classroom levels;
- the evidence and results should be replicated in different samples of the target population, confirming consistent functioning of the M-SE, M-SC and M-ANX measures according to theoretical expectations; and
- the evidence should support directional relationships of M-SC and M-ANX as precursors to M-SE, with all three explaining sufficient variance in mathematics achievement, as suggested by the theoretical and empirical literature on the constructs.

Models and psychometric analysis. The logic guiding each stage of the validation with details of questions and analysis now follows. Each iteration was progressive, in that we evaluated malfunctioning aspects of the instrument, scales or items, discarding or revising the same based on results and generating improved versions for the next iteration.

Iteration 1. Validation of content relevance and representativeness of item pool

Question. Based on objective reviews, to what extent is the content of the originally specified domains, subdomain indicators and items tapping M-SE, M-SC and M-ANX, consistent with the theory and research literature on the constructs?

Procedures and analyses. We invited an expert panel of researchers, measurement experts, teachers and school administrators to evaluate the match of the domains, subdomains and items against theory, with attention to content relevance and content representativeness. As needed, changes were incorporated by adding/deleting items, modifying the language or moving items across different domains.

Iteration 2. Exploratory validation of internal structure and reliability estimates

Questions. Based on an exploratory factor analysis (EFA) of content-validated items from Iteration 1, to what degree do empirically derived factors correspond with item sets tied to M-SE, M-SC and M-ANX domains/subdomains? Are the inter-factor correlations consistent with the literature? What is the internal consistency reliability of factor-defined item sets (scale scores)?

Data. We split the cases in the Year 1 data set randomly to yield separate validation and cross-validation samples for the study. Data Set 1 ($NYr_{1\text{-}1}$ = 255) was used for exploratory analyses and Data Set 2 ($NYr_{1\text{-}2}$ = 251) for confirmatory analysis on the internal structure, convergent validity and internal consistency reliability investigations. No significant differences were obtained in the distributions on background characteristics of the random subsamples shown in Appendix 2.

Analyses. All analyses at this stage were conducted with SPSS Version 19. As factors were expected to be inter-correlated, we used a principal axis factor extraction technique followed by promax rotation of factors. Factors were identified based on eigenvalues > 1, observed breaks in the scree plot and cumulative per cent variance explained. Pattern coefficients (item-to-factor loadings) were examined along with structure coefficients. Items salient to a factor were identified based on minimum loadings of 0.30.

With promax rotation, a pattern coefficient represents the unique relationship between a factor and item/variable, controlling for the other factors (similar in interpretation to standardized regression coefficients). The structure coefficient represents the zero-order correlation between the factor and the variable, and it is often larger in value (Gorsuch, 1983). We followed the standard recommendation to report both (Bandalos and Finney, 2010).

Items that loaded on more than one factor with pattern coefficients approaching 0.30, were treated as ambiguous and dropped. Factors defined by less than 3 items were also not interpreted, with items discarded. Inter-factor correlations were examined to evaluate degrees and direction of convergence. Finally, Cronbach's alpha estimates were obtained for factor-defined scale scores. The results were interpreted against the initial domain framework and literature on M-SE, M-SC and M-ANX.

Iteration 3. Confirmatory analysis of internal structure, convergent validity and reliability

Questions. In a new data set ($N_{Yr1\text{-}2}$ = 251) and using validated items from Iteration 2, to what extent do the data fit a theoretically specified, confirmatory four-factor model, as compared with a single-factor, atheoretical model serving as the baseline? Are the item-to-factor loadings and inter-factor correlations consistent with the literature? What is the reliability of the validated item sets (scale scores)?

Analyses. As the fifth factor (M-SE, *Self-Regulatory Efficacy*) had depressed reliability in data from the immediately preceding iteration, we tested the fit of a four-factor model by dropping the items on that scale. We used Data Set 2 from Year 1 ($N_{Yr1\text{-}2}$ = 251) to investigate the validity of the four-factor confirmatory model representing the M-SE (1), M-SC (2) and M-ANX (1) scales in Iteration 3. For a better evaluation of data-to-model fit, we compared fit of the theoretically specified CFA model against a competing, but atheoretical baseline model with variance of all 25 items explained by one factor.

Factors were permitted to covary. Error terms were hypothesized to be uncorrelated. In each model, the first item loading was constrained to 1.0 to set the scale of measurement, and no items were allowed to load on more than one factor.

Because of the three-point response scales, we treated the item responses as ordinal. With ordinal data, a more robust parameter estimation approach is the weighted least squares (WLS) method, as non-normality is expected in the data (Asparouhov and Muthén, 2009; Flora and Curran, 2004; Muthén *et al.*, 1997). Maximum likelihood (ML) estimation, the more common approach, may not yield accurate estimates, as it assumes multivariate normality and interval level data (Curran *et al.*, 1996; Gold *et al.*, 2003). WLS uses a polychoric correlation matrix to estimate parameters using the inverse of the asymptotic covariance matrix as the weight matrix (Jöreskog, 1990; Muthén, 1984). CFA models were estimated using the robust WLS in the MPlus Version 6 software, or as a weighted least square mean-and variance-adjusted (WLSMV) χ^2 test statistic estimation. WLSMV tends to perform better than WLS, yielding less-biased χ^2 and standard errors (Asparouhov and Muthén, 2009; Flora and Curran, 2004).

Fit was determined based on multiple indices. These included chi-square/df, corrected Bentler comparative fit index (CFI), corrected Tucker–Lewis Index (TLI) and root mean square error of approximation (RMSEA). After Hu and Bentler (1999), we used a cutoff value close to 0.95 or higher on the corrected CFI and TLI indices and close to or below 0.05 on the RMSEA to conclude that a relatively good fit existed between the specified model and the data (Jackson *et al.*, 2009). We also examined results of statistical significance tests for factor loadings, R^2 values, residual and normalized residual matrices and modification indices. Standardized path coefficients (loadings) and inter-factor correlations were interpreted based on consistency with the literature.

Iteration 4. Relationships of measures predicted by Sources of Self-efficacy theory

Questions. Measurement model. Based on the path coefficients, is M-SC represented best with a first- or second-order factor structure? What per cent of variance in M-SE is explained with the latent variables M-SC and M-ANX, treated as Sources of Self Efficacy? Does a causal path model based on Sources of Self-Efficacy theory, with M-SC and M-ANX serving as precursors to M-SE, show acceptable levels of fit and pathways of influence?

Questions. Structural model. Controlling for gender, ethnicity and poverty levels of students, does a causal path model specifying M-SC and M-ANX as precursors to M-SE, and with M-SE having a direct effect on mathematics achievement, show acceptable levels of fit? What per cent of total variance in math achievement is explained by M-SE, M-SC and M-ANX, with gender, ethnicity and poverty levels of students controlled?

Analyses. We proceeded with the four validated factors from the previous iteration and corresponding item sets for the SEM, namely, M-SC, *Positive Affect*; M-SC, *Perceived Competence*; M-ANX, *Negative Affect*; and M-SE, *Perceived Confidence*. The analytic sample for last iteration at the student level included all 469 students from Year 2 (Appendix 2, Table AIII).

Demographic variables included in the analyses as controls were poverty (membership in free/reduced lunch program in school), coded as 1 and 0; race/ethnicity (black versus other), and gender, also coded as 1 and 0 (male versus female). These were selected based on the well-established literature indicating their relationship to math outcomes (Gainor and Lent, 1998; Hampton and Mason, 2003; Usher and Pajares, 2008), to validate directional pathways leading to mathematics achievement.

The achievement outcome measures in the SEM analyses was the total scale score from the New York State standardized mathematics battery. It was administered in the schools late in the spring semester per the state schedule, and it covered the content domains of Number Sense and Operations, Geometry and Algebra. Raw scores for Grades 5-6 had internal consistency reliability estimates of 0.90 and 0.91. The IRT-based scale scores that we used are vertically equated across grades (New York State Education Department, 2007).

Again, given the categorical data, we ran the SEM models using the robust, WLS option in MPlus Version 7.0 (Asparouhov and Muthén, 2009; Flora and Curran, 2004; Muthén *et al.*, 1997). Data to model fit was determined based on a range of fit indices: corrected CFI, corrected TLI and RMSEA with published thresholds for fit (Jackson *et al.*, 2009). We tested the measurement model followed by the structural equation model, based on the Sources of Self-efficacy literature.

Iteration 5. Justifying classroom-level, grouped score-based inferences: descriptive statistics, convergent validity and reliability

Questions. To support interpretations of M-SE, M-SC and M-ANX measures as group score aggregates at classrooms levels, are inter-correlations among mean scores of M-SE, M-SC and M-ANX consistent with underlying theory? Are group scores reliable? Is the variability and distribution of classroom mean scores reasonable?

Analyses. The analytic sample for last iteration included 33 classrooms with 469 students nested within (Appendix 2, Table AIII). We calculated descriptive statistics using classroom mean scores of M-SE, M-SC and M-ANX, and we examined convergent validity levels with Pearson correlations at aggregated levels. To estimate reliability, we obtained the reliability of random coefficients for two-level models using HLM 7 software. Specifically, we ran two-level fully unconditional models, with students nested in classrooms and the M-SE, M-SC and M-ANX classroom means (random coefficients) at Level 2 (for HLM definitions of reliability, see: www.ssicentral.com/hlm/help6/faq/Reliability_of_random_coefficients.pdf).

Phase V. Evidence evaluation with reference to Phase I, assessment context specifications. The final phase involved a comprehensive evaluation of the all the evidence for the proposed student level inferences and uses, as well as for classroom-level inferences and uses. This is presented under the Discussion section, following the Results.

Results

Content validation results from Iteration 1

The updated literature review by a member of the research team who had not participated in the initial instrument design and pilot-testing phase, verified the content-based validity of indicators of the domains. Only Item 5 was reclassified and placed under M-SC, instead of under M-SE. Following further content validation and minor language revisions to items by external experts, the item breakdown by construct was as follows:

- *M-SE*: 20 items organized in two subdomains (*Perceived Confidence* and *Self-Regulatory Efficacy* in mathematics).
- *M-SC*: 11 items organized in two subdomains (*Perceived Competence* and *Positive Affect* in mathematics).
- *M-ANX*: 6 items organized in a single domain, with 3 incorporating reverse-oriented wording (*Negative Affect* in mathematics).

Validity evidence from Iteration 2

Internal structure. See Table I and Appendix 1. The initial EFA with all 37 items produced a scree plot showed a slight bend after five factors. Cumulative percentage variance explained by the first five factors was 44 per cent. The promax rotated five-factor solution rendered 25

Factor		Items retained
1	M-SC, *Positive Affect*	22, 23, 29, 30, 31, 33, 35, 37
2	M-SC, *Perceived Competence*	5, 8, 25, 26, 27, 28
3	M-ANX, *Negative Affect*	32, 34, 36
4	M-SE, *Perceived Confidence*	1, 2, 3, 4
5	M-SE, *Self-regulatory Efficacy*	12, 19, 20, 21

Note: Factor 5 was defined by items 10-14 in cross-validation sample

Table I. Evidence of internal structure: Factor defined scales ($N_{Yr1\text{-}1} = 255$)

items that loaded on factors corresponding with the specified dimensions and suggested a five-scale structure.

Factors 1-2 corresponded with the theoretically specified dimensions of M-SC: *Positive Affect* and *Perceived Competence* in math, with loadings of 0.30-0.79 and 0.38-0.88, respectively. Factors 4 and 5 matched two dimensions specified for the M-SE construct, namely, *Perceived Confidence* and *Self-Regulatory Efficacy*, with loadings of 0.44-0.68 and 0.37-0.58, respectively. Factor 3 was represented by three of six negatively oriented items of M-ANX, with loadings of 0.63 and 0.76, respectively. All loadings were above the 0.30 cutoff, with a depressed loading of 0.24 on Item 4 ruled out, given the loadings of 0.44 or higher in cross-validation samples on that item.

The three positively worded items under M-ANX loaded on the *Positive Affect* dimension of M-SC instead. This result verified our earlier suspicion that positive emotional language in M-ANX items may lead to responses denoting either M-SC or M-SE. We found that overlap empirically with M-SC.

In sum, the initial EFA results suggested a five-scale structure where M-SE and M-SC had two dimensions each, consistent with the theoretical framework underlying the constructs. M-ANX was also validated as consistent with existing literature and reflective of negative math affect (Bandura, 1986, 1997; Bong and Skaalvik, 2003; Marsh, 1999; Pajares and Schunk, 2002; Usher and Pajares, 2008, 2009).

However, 12 items were dropped from the 37-item survey as factor loadings fell below the threshold set, or items loaded ambiguously on multiple factors without yielding a simple structure. These high loadings suggested a lack of differentiation in the constructs we investigated in the Grade 5 (Year 1) sample, making it difficult to derive clean measures with sufficiently distinct information from the younger respondents in the sample.

Internal consistency reliability. Table II provides descriptive statistics on scale scores. The items defined by each of the five factors were tested for internal consistency reliability in the Year 1 subsample and cross-validated again in Year 2. Cronbach's α estimates ranged from 0.85 to 0.70. Two exceptions were α value of 0.55 (the *Self-Regulatory* dimension of M-SE) which dropped to 0.34 in the cross-validation sample, and α value at 0.62 for M-ANX, which improved in the cross-validation sample to >0.70.

Convergent validity. The inter-factor correlations in Table III suggested consistency with theoretical expectations and were cross-validated in four of the five factor-defined scales. *Perceived Competence* factor and *Positive Affect* factor of M-SC correlated at 0.41; the *Perceived Confidence* factor of M-SE correlated positively at 0.28 and that of M-SC factors at 0.53; and M-ANX factor was uncorrelated or correlated negatively with all of the above at −0.43, 0.02 and −0.18. The *Self-Regulatory* dimension of M-SE correlated at 0.19 and 0.34 with M-SC factors, and was −0.10 with M-ANX. This suggested theoretical meaningfulness of all the factors.

Table II.
Descriptive statistics of scale scores at the student level

			Student level ($N_{Yr1\text{-}1}$ = 255)			
Factor-defined Scales	Mean	SD	Minimum	Maximum	Range	Reliability
M-SC, *Positive Affect*	16.33	4.47	8	24	8-24	0.85
M-SC, Perceived competence	14.76	3.12	6	18	6-18	0.85
M-ANX, *Negative Affect*	4.43	1.76	3	9	3-9	0.62
M-SE, *Perceived Confidence*	10.96	1.34	4	12	4-12	0.70
M-SE, *Self-regulatory Efficacy*	10.53	1.57	4	12	4-12	0.55

Validity evidence from Iteration 3

See Figure 2. The depressed reliability of the fifth scale at 0.55 (M-SE, *Self-Regulatory Efficacy*) needed further examination. Therefore, we proceeded with four evidence-supported scales, keeping the fifth on stand-by for further checking and evaluation with new data.

In a comparative evaluation of the four-factor CFA results against a baseline model where all 25 items were explained by a single latent factor (not shown), the latter did not show good fit: χ^2 = 1,041.51, df = 275, CFI = 0.83, TLI = 0.81, RMSEA = 0.11. In comparison, the four-factor CFA model in Figure 2 yielded far better fit (χ^2 = 351.09, df = 183, CFI = 0.96, TLI = 0.96, RMSEA = 0.06), meeting published thresholds for good fit on the CFI and TLI, with RMSEA just above the 0.05 cut-off.

We also found high levels of fit with a five-factor model (χ^2 = 424.57, df = 265, CFI = 0.96, TLI = 0.96, RMSEA = 0.05; not shown). However, the previously-reported reliability issues and unstable item-factor composition of the M-SE *Self-Regulatory* scale led to a decision to re-examine the language and content of items before further validation.

Convergent validity. See Figure 2. The standardized item-to-latent factor loadings of the four-factor CFA model were also acceptable, lying predominantly between 0.60 and 0.80. Standard errors were low to very low. The inter-correlations are all consistent with theory, as follows:

- Factor 1 (M-SC, *Positive Affect*) and Factor 3 (M-ANX, *Negative affect*) correlate at −0.109.
- Factor 2 (M-SC, *Perceived Competence*) and Factor 3 (M-ANX) correlate at −0.526.
- Factor 4 (M-SE, *Perceived Confidence*) correlates with Factor 3 (M-ANX) at −0.50.
- Factors 1, 2 and 4 inter-correlate positively and robustly at 0.72, 0.39 and 0.63, respectively.

Very similar levels of correlation among M-SE, M-SC and M-ANX construct measures are reported in the literature discussed.

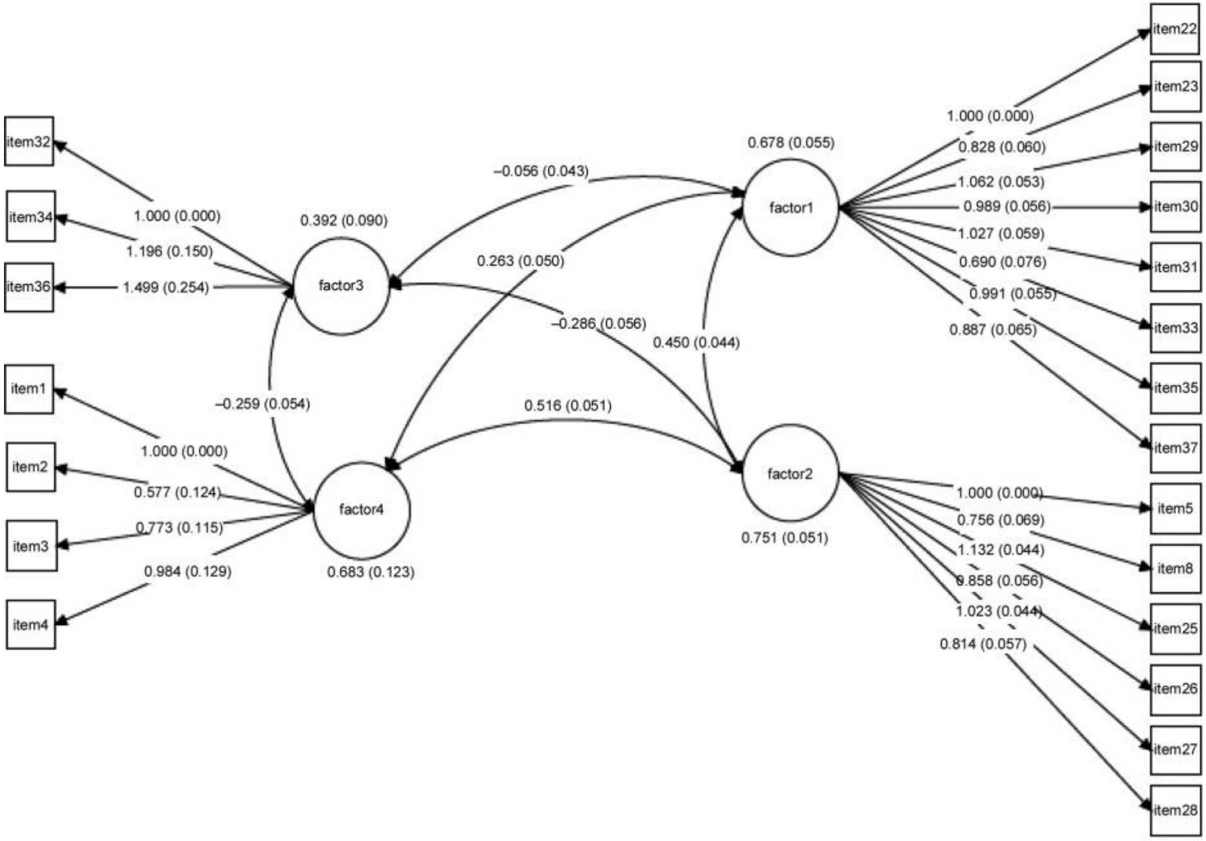
Reliability. Cronbach's α reliability estimates for scale scores were at 0.83 (*affective* dimension of M-SC), 0.84 (*Perceived Competence* dimension of M-SC), 0.67 (*Perceived Competence* dimension of M-SE) and 0.77 (M-ANX).

Validity evidence from Iteration 4

Both first- and second-order measurement models for M-SC, with M-ANX and M-SE built in (figures not shown; available upon request), were found to have comparable levels of fit (χ^2 = 438.72, df = 183, CFI = 0.96, TLI = 0.96, RMSEA = 0.05 for the first-order structure; χ^2 = 450.46, df = 184, CFI = 0.96, TLI = 0.96, RMSEA = 0.05 for the second-order structure). However, with M-SC modeled with a second-order factor structure, the path coefficients provided stronger confirmation of a majority of our hypotheses about the construct measures.

	Student level (N_{Yr1-1} = 255)				
Construct Measures	(1)	(2)	(3)	(4)	(5)
1. M-SC, *Positive Affect*	1.00				
2. M-SC, *Perceived Competence*	0.41	1.00			
3. M-ANX, *Negative Affect*	0.02	−0.43	1.00		
4. M-SE, *Perceived Confidence*	0.28	0.53	−0.18	1.00	
5. M-SE, *Self-regulatory Efficacy*	0.19	0.34	−0.10	0.30	1.00

Table III. Inter-factor correlations at the student-level *(N_{Yr1-1} = 255)*



Notes: $\chi^2 = 351.09$, df = 183, CFI = 0.96, TLI = 0.96, RMSEA = 0.06; Factor 1 = M-SC, *Positive Affect*; Factor 2 = M-SC, *Perceived Competence*; Factor 3 = M-ANX, *Negative Affect*; and Factor 4 = M-SE, *Perceived Confidence*

Figure 2.
Four-factor confirmatory factor model ($N_{Yr1\text{-}2} = 251$)

Results of the directional measurement model confirmed convergent validity of the four non-cognitive factors based on path coefficients. The estimated R^2 value was 0.587 (SE = 0.07), indicating that 59 per cent of the variance in *Perceived Confidence* factor of M-SE was explained with M-SC and M-ANX as antecedent factors. M-ANX influenced M-SE directionally at −0.299 (SE = 0.06), as predicted. M-SC influenced M-SE directionally at +0.590 (SE = 0.06), as predicted. The two Sources, M-SC with a hierarchical factor structure and M-ANX, inter-correlated at −0.426 (SE = 0.06) as predicted.

Figure 3 shows the directional path model showing the direct and indirect effects of non-cognitive factors with students' mathematics achievement serving as an observed, outcome variable. Again, we obtained high levels fit for the model. Fit was estimated at: χ^2 = 575.33, df = 265, CFI = 0.96, TLI = 0.95, RMSEA = 0.05. Variance explained in achievement is in the order of 34 per cent, controlling for ethnicity, gender and poverty. The direct effects of the M-SE, *Perceived Confidence* factor on achievement is 0.242 (SE = 0.08). As compared to the influences of M-SC at 0.195 (SE.07) and M-ANX at −0.184 (SE = 0.06), it presents the strongest estimated effect. All three effects are statistically significant at or under the 10 per cent error level.

Mediated by M-SE, the indirect effects of M-SC on achievement were estimated at 0.133 and those of M-ANX at 0.076 (not shown). A comparison of the estimated direct and negative directional effect of M-ANX on achievement (−0.184), versus its indirect effect of near zero (0.076), suggested a mitigating effect of M-SE as a mediator. This finding is affirmed by the literature surveyed on Sources of Self efficacy, and it is particularly encouraging from a construct validity standpoint at the student level.

Validity evidence from Iteration 5

See Tables IV and V. We see that with the small sample of 33 classrooms, the variability of means is somewhat restricted in three of four construct measures. Only the M-SC, *Positive Affect* scale mean has a SD of 2.05, with others at <1.0. The estimated reliability of random coefficients was affected by this limited variance. It was 0.725 on M-SC (*Positive Affect*, Factor 1), with estimates falling under 0.50 for the remaining three constructs at the classroom level.

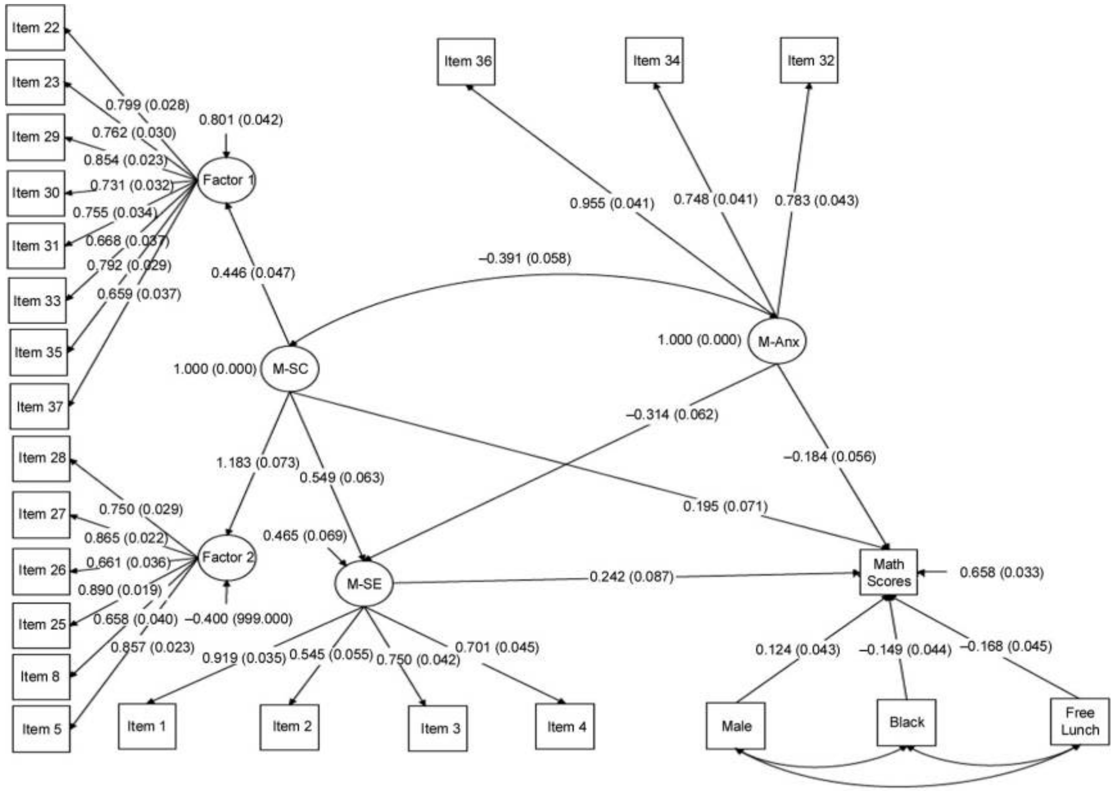
Encouragingly, the convergent validity coefficients suggested theoretical meaningfulness of the four constructs aggregated at the classroom level: M-SE (*Perceived Confidence*, Factor 4) correlated at 0.18 and 0.50 with M-SC scale means; M-SC (*Positive Affect*, Factor 1) correlated at 0.43 with (*Perceived Competence*, Factor 2). M-ANX correlated with M-SE (*Perceived Confidence*), M-SC (*Positive Affect*) and M-SC (*Perceived Competence*) at −0.08, −11 and −0.32, respectively.

Discussion: evidence appraisal on M-SC, M-ANX and M-SE measures

Student-level inferences and uses from measures

Cycling back to the intended inferences and uses specified in Phase I for the individual student-level measures, we see that the evidence taken together supports inferences tied to the underlying M-SE, M-SC and M-ANX constructs on four scales (Tables I-V, Figures 2). These scales yielded parsimonious, individually informative and theoretically meaningful measures. The validation of the Sources of Self-efficacy theory and support for directional influences of these four validated non-cognitive constructs on mathematics achievement (Figure 3) build further confidence in the current measures for the inferences and uses specified. Each scale is yielding data with satisfactory levels of unique variance, and the results were cross-validated and replicated in independent samples of students from Years 1-2.

The evidence from directional path models on the M-SE *Perceived Confidence* factor, and its consistency with the Sources of Self Efficacy literature, merits some discussion (Usher



Notes: Model controls for student gender, race/ethnicity, and poverty status. $\chi^2 = 541.29$, df = 265, CFI = 0.96, TLI = 0.96, RMSEA = 0.05. R^2 for Math Score = 0.34 (S.E. = 0.03). Numbers shown are the standardized regression coefficients. All coefficients are statistically significant at 0.01 level. Factor 1 = M-SC, *Positive Affect*; Factor 2 = M-SC, *Perceived Competence*

Figure 3.
Structural equation model depicting the associations between mathematics achievement and M-SE, M-SC, and M-ANX (N_{Year2} =466)

and Pajares, 2008). Although causality cannot be claimed firmly without experimental controls and manipulated variables, the SEM results confirmed that M-SC and M-ANX together are likely precursors to the *Perceived Confidence* dimension of M-SE, with that dimension of M-SE serving as an effective mediator of students' mathematics achievement. That the non-cognitive factors together explained 34 per cent of the variance in student achievement, and that M-SE diminished the direct negative influence of M-ANX when serving as a mediator, is an important finding for educational research, practice and policy.

The exception, however, was the scale tapping into the *Self-Regulatory Efficacy* subdomain of M-SE. The item inconsistencies and malfunctions on that scale were identified based on inconsistent reliability evaluations, item-factor composition and inter-factor correlations in different data sets, and need to be investigated further.

Why was the evidence mixed for the *Self-Regulatory* M-SE factor? The content of the items must be re-examined. Respondents may have been too young to generate sound survey data on the construct. Because it involves metacognition, it could also be that the *Self-Regulatory* dimension of M-SE is still undifferentiated in pre-adolescents, needing more time to develop. Further research will continue to improve this scale.

Regardless, we conclude that the body of evidence suggests overall construct validity for four measures (of five) supporting individual, student-level interpretations and uses by teachers, researchers and practitioners. The validated scales may also be useful as non-cognitive indicators for monitoring school-based reforms at the student level, for low-stakes actions.

Classroom-level inferences and uses. The convergent validity evidence supports theoretical meaningfulness of the four measures at the classroom level, too. The depressed reliability results may have been an artifact of the relatively small sample for the project. Further validation is necessary with larger samples of classrooms and more variable data sets. Given that the intended inferences and uses with the classroom means were for formative evaluations of the intervention, there were no untoward consequences for individuals and schools involved in the project. However, because the reliability of classroom means was a disappointment with the present data set, that evidence at the classroom level should be treated as preliminary.

	Classroom level ($N_{Yr2\_Classroom}$ = 33)				
Variables	Mean	SD	Minimum	Maximum	Range
M-SC, *Positive Affect*	16.14	2.05	12.38	20.47	12.38-20.47
M-SC, *Perceived Competence*	14.69	0.88	12.77	16.31	12.77-16.31
M-ANX, *Negative Affect*	4.73	0.69	3.44	6.64	3.44-6.64
M-SE, *Perceived Confidence*	10.96	0.41	10.06	11.60	10.06-11.60

Table IV. Descriptive statistics of mean scores at classroom-level

	Classroom level ($N_{Yr2\_Classroom}$ = 33)			
Construct measures	(1)	(2)	(3)	(4)
1. M-SC, *Positive Affect*	1.00			
2. M-SC, *Perceived Competence*	0.43	1.00		
3. M-ANX, *Negative Affect*	−0.11	−0.32	1.00	
4. M-SE, *Perceived Confidence*	0.18	0.50	−0.08	1.00

Table V. Inter-factor correlations at the classroom-level

Assessment user involvement. The teachers, leaders and members of the school district were participating partners in the research project. Assessment users were involved during instrument pilot-tests (Phase III), content validation (Phase IV), data collection (Phase IV) and informal evidence evaluations (Phase V). There were frequent exchanges among assessment designers/researchers and assessment stakeholders in schools. Although eventually the construct measures and data were intended for use in ongoing school improvement efforts, within the parameters of the project, uses of assessment results were all "low stakes".

This advantage helped build ownership levels and improved assessment data uses by teachers over time, even among unionized teachers who were unwilling project participants at first (Chatterji, 2012). This observation is consistent with principles of participatory evaluation on which the process model is founded (Patton, 2008).

Utility of user-centered process

In typical educational measurement practice today, instruments and measures are rarely validated for all population units on whom measure-based inferences are to be drawn, or actions may be taken, such as, students, classrooms, schools and districts. Such issues have plagued the measurement field in general, engendering recent debates among validity theorists. Some even recommend that during validation, how assessment results are used or misused, could be ignored (Borsboom *et al.*, 2004).

By adding pragmatic ideas on users and user-centeredness, the approach we applied might offer solutions for upholding validity principles in applied contexts. The user-centered, iterative methodology (Figure 1) draws on consensus-based, best practices recommended in measurement today (AERA *et al.*, 2014).

Limitations, implications and future research

Directionality of the variable influences in Figure 3 needs continuing verification. The question may be raised as to whether the relationships would be similar if the paths were in the reverse direction. Future research should attempt to replicate results by manipulating time gaps in measuring M-SC and M-ANX (as precursors), M-SE (as mediator) and math achievement (as outcome) in models with similar age samples.

To boost mathematics achievement levels of students, educational researchers and policymakers alike are beginning to recognize the importance of developing non-cognitive capacities of students today (Education Database Online Blog, 2013; Entwisle *et al.*, 2005; Gutman and Schoon, 2013). The evidence from this study endorses that course of action. We hope the validated construct measures will facilitate further research, as well as find utility in practice and policy contexts in mathematics education settings. Research is continuing on the self-regulatory efficacy scale of M-SE.

References

AERA, APA. and NCME (2014), *Standards for educational and psychological testing, American Educational Research Association*, Washington, DC, available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:NCME+(1999)+Standards+for+educational+and+psychological+testing#0

Ahmed, W., Minnaert, A., Kuyper, H. and van der Werf, G. (2012), "Reciprocal relationships between math self-concept and math anxiety", *Learning and Individual Differences*, Vol. 22 No. 3, pp. 385-389, available at: http://doi.org/10.1016/j.lindif.2011.12.004

Ashcraft, M.H. (2001), "The relationships among working memory, math anxiety, and performance", *Journal of Experimental Psychology, General*, Vol. 130 No. 2, pp. 224-237.

Ashcraft, M.H. (2002), "Math anxiety: personal, educational, and cognitive consequences", *Current Directions in Psychological Science*, Vol. 11 No. 5, pp. 181-185, available at: http://doi.org/10.1111/1467-8721.00196

Asparouhov, T. and Muthén, B.O. (2009), "Exploratory structural equation modeling", *Structural Equation Modeling*, Vol. 16 No. 3, pp. 397-438, available at: http://doi.org/10.1080/10705510903008204

Bandalos, D.L. and Finney, S.J. (2010), "Factor analysis: exploratory and confirmatory", in Hancock, G.R. and Mueller, R.O. (Eds), *The Reviewer's Guide to Quantitative Methods in the Social Sciences*, Routledge, Abingdon, pp. 93-114.

Bandalos, D.L., Yates, K. and Thorndike-Christ, T. (1995), "Effects of math self-concept, perceived self-efficacy, and attributions for failure and success on test anxiety", *Journal of Educational Psychology*, Vol. 87 No. 4, pp. 611-623.

Bandura, A. (1986), *Social Foundations of Thought and Action: A Social Cognitive Theory*, Prentice-Hall, Englewood Cliffs, NJ.

Bandura, A. (1993), "Perceived self-efficacy in cognitive development and functioning", *Educational Psychologist*, Vol. 28 No. 2, available at: http://doi.org/10.1207/s15326985ep2802_3

Bandura, A. (1997), *Self-Efficacy: The Exercise of Control*, Freeman, New York, NY, available at: http://doi.org/10.5860/CHOICE.35-1826

Bandura, A. (2000), "Self-efficacy: the foundation of agency", in Perrig, W.J. and Grob, A. (Eds), *Control of Human Behaviour, Mental Processes and Consciousness*, Erlbaum, Mahwak, NJ, pp. 17-33.

Bandura, A. (2006), "Guide for constructing self-efficacy scales", in Pajares, F. and Urdan, T. (Eds), *Self-Efficacy Beliefs of Adolescents*, Information Age Publishing, Greenwich, CT, pp. 307-337.

Bong, M. and Clark, R.E. (1999), "Comparison between self-concept and self-efficacy in academic motivation research", *Educational Psychologist*, Vol. 34 No. 3, pp. 139-153, available at: http://doi.org/10.1207/s15326985ep3403_1

Bong, M. and Skaalvik, E.M. (2003), "Academic self-concept and self-efficacy: how different are they really?", *Educational Psychology Review*, Vol. 15 No. 1, pp. 1-40, available at: http://doi.org/10.1023/A:1021302408382

Borsboom, D., Mellenberh, G.J. and van Heerden, J. (2004), "The concept of validity", *Psychological Review*, Vol. 111 No. 4, p. 1061.

Chatterji, M. (2003), *Designing and Using Tools for Educational Assessment*, Allyn & Bacon/Pearson, Boston, MA.

Chatterji, M. (2012), "Development and validation of indicators of teacher proficiency in diagnostic classroom assessment", *The International Journal of Educational and Psychological Assessment*, Vol. 9 No. 2, pp. 4-25.

Chatterji, M. (*in press*), *Designing Assessments for Multidisciplinary Constructs and Applications: A User-Centered Methodology*, Guilford Publications, New York, NY.

Chatterji, M., Koh, N., Choi, L. and Iyengar, R. (2009), "Closing learner gaps proximally with teacher-mediated diagnostic assessment", *Research in the Schools*, Vol. 16 No. 2, pp. 60-77.

Cronbach, L.J. (1971), "Test validation", in Thorndike, R.L. (Ed), *Educational Measurement* (2nd ed.), American Council on Education, Washington DC, p. 443.

Cronbach, L.J. and Meehl, P.E. (1955), "Construct validity in psychological tests", *Psychological Bulletin*, Vol. 52 No. 4, pp. 281-302, available at: http://doi.org/10.1037/h0040957

Curran, P.J., West, S.G. and Finch, J.F. (1996), "The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis", *Psychological Methods*, Vol. 1 No. 1, pp. 16-29, available at: http://doi.org/10.1037/1082-989X.1.1.16

DeVellis, R.F. (2003), *Scale Development: Theory and Applications*, SAGE publication, Thousand Oaks, CA.

Duckor, B. (2017), "Got grit? Maybe. . .", *Phi Delta Kappan*, Vol. 98 No. 7, pp. 61-66.

Eccles, J.S. (2005), "Subjective task value and the Eccles *et al.* model of achievement-related choices", in Elliot, A.J. and Dweck, C.S. (Eds), *Handbook of Competence and Motivation*, Guilford, New York, NY, pp. 105-121.

Education Database Online Blog (2013), "*Noncognitive measures: The academic trend that could change everything*", available at: www.onlineeducation.net/2013/02/04/noncognitive-measures-the-academic-trend-that-could-change-everything

Entwisle, D.R., Alexander, K.L. and Olson, L.S. (2005), "First grade and educational attainment by age 22: a new story", *American Journal of Sociology*, Vol. 110 No. 5, pp. 1458-1502, available at: http://doi.org/10.1086/428444

Ferla, J., Valcke, M. and Cai, Y. (2009), "Academic self-efficacy and academic self-concept: reconsidering structural relationships", *Learning and Individual Differences*, Vol. 19 No. 4, pp. 499-505, available at: http://doi.org/10.1016/j.lindif.2009.05.004

Flora, D.B. and Curran, P.J. (2004), "An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data", *Psychological Methods*, Vol. 9 No. 4, pp. 466-491, available at: http://doi.org/10.1037/1082-989X.9.4.466

Gainor, K.A. and Lent, R.W. (1998), "Social cognitive expectations and racial identity attitudes in predicting the math choice intentions of black college students", *Journal of Counseling Psychology*, Vol. 45 No. 4, pp. 403-413, available at: http://doi.org/10.1037/0022-0167.45.4.403

Galla, B.M. and Wood, J.J. (2012), "Emotional self-efficacy moderates anxiety-related impairments in math performance in elementary school-age youth", *Personality and Individual Differences*, Vol. 52 No. 2, pp. 118-122, available at: http://doi.org/10.1016/j.paid.2011.09.012

Gold, M.S., Bentler, P.M. and Kim, K.H. (2003), "A comparison of maximum-likelihood and asymptotically distribution-free methods of treating incomplete nonnormal data", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 10 No. 1, pp. 47-79, available at: http://doi.org/10.1207/S15328007SEM1001_3

Gorsuch, R.L. (1983), *Factor Analysis*, 2nd ed., Lawrence Erlbaum, Hillsdale, NJ, available at: http://doi.org/10.1108/EUM0000000002688

Gutman, L.M. and Schoon, I. (2013), *The impact of non-cognitive skills on outcomes for young people: Literature review*, Leading education and social research.

Hampton, N.Z. and Mason, E. (2003), "Learning disabilities, gender, sources of efficacy, self-efficacy beliefs, and academic achievement in high school students", *Journal of School Psychology*, No. 41, pp. 101-112, available at: http://doi.org/10.1016/S0022-4405(03)00028-1

Hembree, R. (1990), "The nature, effects, and relief of mathematics anxiety", *Journal for Research in Mathematics Education*, Vol. 21 No. 1, pp. 33-46, available at: http://doi.org/10.2307/749455

Hoffman, B. (2010), "I think I can, but I'm afraid to try": the role of self-efficacy beliefs and mathematics anxiety in mathematics problem-solving efficiency", *Learning and Individual Differences*, Vol. 20 No. 3, pp. 276-283, available at: http://doi.org/10.1016/j.lindif.2010.02.001

Hu, L. and Bentler, P.M. (1999), "Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 6 No. 1, pp. 1-55, available at: http://doi.org/10.1080/10705519909540118

Jackson, D.L., Gillaspy, J.A. and Purc-Stephenson, R. (2009), "Reporting practices in confirmatory factor analysis: an overview and some recommendations", *Psychological Methods*, Vol. 14 No. 1, pp. 6-23, available at: http://doi.org/10.1037/a0014694

Jain, S. and Dowson, M. (2009), "Mathematics anxiety as a function of multidimensional self-regulation and self-efficacy", *Contemporary Educational Psychology*, Vol. 34 No. 3, pp. 240-249, available at: http://doi.org/10.1016/j.cedpsych.2009.05.004

Joët, G., Usher, E.L. and Bressoux, P. (2011), "Sources of self-efficacy: an investigation of elementary school students in France", *Journal of Educational Psychology*, Vol. 103 No. 3, pp. 649-663, available at: http://doi.org/10.1037/a0024048

Jöreskog, K.G. (1990), "New developments in LISREL: analysis of ordinal variables using polychoric correlations and weighted least squares", *Quality and Quantity*, Vol. 24 No. 4, pp. 387-404.

Kane, M.T. (2006), "Validation", in Brennan, R.L. (Ed.), *Educational Measurement*, 4th ed., Praeger Publishers, Westport, CT, pp. 17-64.

Kane, M.T. (2013), "Validating the interpretations and uses of test scores", *Journal of Educational Measurement*, Vol. 50 No. 1, pp. 1-73.

Lee, J. (2009), "Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries", *Learning and Individual Differences*, Vol. 19 No. 3, pp. 355-365, available at: http://doi.org/10.1016/j.lindif.2008.10.009

Marsh, H.W. (1999), "Academic self description questionnaire-I: ASDQ I, *University of Western Sydney, Self-Concept Enhancement and Learning Facilitation Research Centre*", Macarthur, available at: http://doi.org/10.1037/0022-0663.82.4.623

Marsh, H.W. and Martin, A.J. (2011), "Academic self-concept and academic achievement: relations and causal ordering", *British Journal of Educational Psychology*, Vol. 81 No. 1, pp. 59-77, available at: http://doi.org/10.1348/000709910X503501

Marsh, H.W., Roche, L.A., Pajares, F. and Miller, D. (1997), "Item-specific efficacy judgments in mathematical problem solving: the downside of standing too close to trees in a forest", *Contemporary Educational Psychology*, Vol. 22 No. 3, pp. 363-377, available at: http://doi.org/10.1006/ceps.1997.0942

Messick, S. (1989), "Meaning and values in test validation: the science and ethics of assessment", *Educational Researcher*, Vol. 18 No. 2, pp. 5-11.

Mislevy, R.J. (2006), "Cognitive psychology and educational assessment", in Brennan, R.L. (Ed.), *Educational Measurement* (4th ed.), Praeger Publishers, Westport, CT, pp. 257-306.

Mullen, C.A. and Schunk, D.H. (2011), "The role of the professional learning community in dropout prevention", *AASA Journal of Scholarship & Practice*, Vol. 8 No. 3, pp. 26-29, available at: http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ944702&site=ehost-live%5Cn, http://www.aasa.org/jsp.aspx

Muthén, B.O. (1984), "A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators", *Psychometrika*, Vol. 49 No. 1, pp. 115-132, available at: http://doi.org/10.1007/BF02294210

Muthén, B.O., Du Toit, S.H.C. and Spisic, D. (1997), "Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes", available at: http://doi.org/10.2139/ssrn.201668

New York State Education Department (2007), *New York State Testing Program 2007: Mathematics Grade 3-8*, McGraw-Hill, New York, NY.

Pajares, F. (1997), "Current directions in self-efficacy research", in Maehr, M. and Pintrich, P.R. (Eds), *Advances in Motivation and Achievement*, JAI Press, Greenwich, CT, pp. 1-49.

Pajares, F. (2002), "Gender and perceived self-efficacy in self-regulated learning", *Theory into Practice*, Vol. 41 No. 2, pp. 116-125, available at: http://doi.org/10.1207/s15430421tip4102_8

Pajares, F. and Graham, L. (1999), "Self-efficacy, motivation constructs, and mathematics performance of entering middle school students", *Contemporary Educational Psychology*, Vol. 24 No. 2, pp. 124-139, available at: http://doi.org/10.1006/ceps.1998.0991

Pajares, F. and Miller, M.D. (1994), "Role of self-efficacy and self-concept beliefs in mathematical problem solving: a path analysis", *Journal of Educational Psychology*, Vol. 86 No. 2, pp. 193-203, available at: http://doi.org/10.1037/0022-0663.86.2.193

Pajares, F. and Schunk, D.H. (2002), "Self and self-belief in psychology and education: a historical perspective", *Improving Academic Achievement: Impact of Psychological*

Factors on Education, Vol. 27 No. 765, pp. 3-21, available at: http://doi.org/10.1016/B978-012064455-1/50004-X

Patton, M. (2008), *Utilization-Focused Evaluation*, Sage Publications, Thousand Oaks, CA, available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Utilization-Focused+Evaluation:+The+New+Century+Text#0

Perry, J.C., DeWine, D.B., Duffy, R.D. and Vance, K.S. (2007), "The academic self-efficacy of urban youth-a mixed-methods study of a school-to-work program", *Journal of Career Development*, Vol. 34 No. 2, pp. 103-126, available at: http://doi.org/10.1177/0894845307307470

Pietsch, J., Walker, R. and Chapman, E. (2003), "The relationship among self-concept, self-efficacy, and performance in mathematics during secondary school", *Journal of Educational Psychology*, Vol. 95 No. 3, pp. 589-603, available at: http://doi.org/10.1037/0022-0663.95.3.589

Porter, A.C., McMaken, J., Hwang, J. and Yang, R. (2011), "Common core standards: the new US intended curriculum", *Educational Researcher*, Vol. 40 No. 3, pp. 103-116, available at: http://doi.org/10.3102/0013189X11424697

Ravitch, D. (2011), "*School "reform": A failing grade*", available at: www.nybooks.com/articles/2011/09/29/school-reform-failing-grade/?pagination=false

Roeser, R.W., Eccles, J.S. and Sameroff, A.J. (2000), "School as a context of early adolescents' academic and social-emotional development: a summary of research findings", *The Elementary School Journal*, Vol. 100 No. 5, pp. 443-471.

Schunk, D.H. and Usher, E.L. (2011), "Assessing self-efficacy for self-regulated learning", *Handbook of Self-Regulation of Learning and Performance*, Routledge, New York, NY, pp. 282-297, available at: http://doi.org/10.1037/t21623-000

Seaton, M., Parker, P., Marsh, H.W., Craven, R.G. and Yeung, A.S. (2014), "The reciprocal relations between self-concept, motivation and achievement: juxtaposing academic self-concept and achievement goal orientations for mathematics success", *Educational Psychology*, Vol. 34 No. 1, pp. 49-72, available at: http://doi.org/10.1080/01443410.2013.825232

Usher, E.L. and Pajares, F. (2006), "Sources of academic and self-regulatory efficacy beliefs of entering middle school students", *Contemporary Educational Psychology*, Vol. 31 No. 2, pp. 125-141, available at: http://doi.org/10.1016/j.cedpsych.2005.03.002

Usher, E.L. and Pajares, F. (2008), "Sources of self-efficacy in school: critical review of the literature and future directions", *Review of Educational Research*, Vol. 78 No. 4, pp. 751-796, available at: http://doi.org/10.3102/0034654308321456

Usher, E.L. and Pajares, F. (2009), "Sources of middle school students' self-efficacy in mathematics: a validation study", *Contemporary Educational Psychology*, Vol. 34, pp. 89-101, available at: http://doi.org/10.3102/0002831208324517

Vick, R.M. and Packard, B.W.-L. (2008), "Academic success strategy use among community-active urban Hispanic adolescents", *Hispanic Journal of Behavioral Sciences*, Vol. 30 No. 4, pp. 463-480, available at: http://doi.org/10.1177/0739986308322913

Vrugt, A. (2004), "Perceived self-efficacy and work motivation", *Advances in Psychology Research*, Vol. 31 No. 8, pp. 147-171.

Wang, Z., Osterlind, S.J. and Bergin, D.A. (2012), "Building mathematics achievement models in four countries using TIMSS 2003", *International Journal of Science and Mathematics Education*, Vol. 10 No. 5, pp. 1215-1242, available at: http://doi.org/10.1007/s10763-011-9328-6

Wigfield, A. and Eccles, J.S. (2000), "Expectancy-value theory of achievement motivation", *Contemporary Educational Psychology*, Vol. 25 No. 1, pp. 68-81, available at: http://doi.org/10.1006/ceps.1999.1015

Zimmerman, B.J. (1989), "A social cognitive view of self-regulated academic learning", *Journal of Educational Psychology*, Vol. 81 No. 3, pp. 329-339, available at: http://doi.org/10.1037/0022-0663.81.3.329

Zimmerman, B.J. (1994), "Dimensions of academic self-regulation: a conceptual framework for education", *Self-Regulation of Learning and Performance: Issues and Educational Applications*, Vol. 1, pp. 3-21.

Zimmerman, B.J. (2000), "Self-efficacy: an essential motive to learn", *Contemporary Educational Psychology*, Vol. 25 No. 1, pp. 82-91, available at: http://doi.org/10.1006/ceps.1999.1016

Zimmerman, B.J. and Bandura, A. (1994), "Impact of self-regulatory influences on writing course attainment", *American Educational Research Journal*, Vol. 31 No. 4, available at: http://doi.org/10.3102/00028312031004845

Zimmerman, B.J. and Martinez-Pons, M. (1990), "Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use", *Journal of Educational Psychology*, Vol. 82 No. 1, pp. 51-59, available at: http://doi.org/10.1037/0022-0663.82.1.51

Further reading

Chatterji, M. and Lin, M. (2016), *Validating Relationships of Mathematics-related Self-efficacy, Self-concept, and Anxiety with Achievement*, Paper presentation at the 2016 National Council on Measurement in Education Annual Meeting, Washington, DC.

Lin, M. and Chatterji, M. (2012), *Survey-based Non-cognitive Measures for Young Respondents: Tackling Errors Using a Multi-stage Validation Approach*, Paper presentation at the 2012 National Council on Measurement in Education Annual Meeting at Vancouver, Canada.

Appendix 1

Table AI.
Items organized by construct and results of exploratory factor analysis (EFA)

	Construct and items	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Mathematics Self-Efficacy (M-SE) items						
1	I can do the math work that my teacher gives me	0.07 (0.27)	0.00 (0.40)	0.02 (−0.14)	_0.68* (0.71)_	0.03 (0.20)
2	I can do the math problems in my textbook	0.04 (0.18)	−0.03 (0.25)	0.08 (−0.05)	_0.37* (0.44)_	−0.14 (0.08)
3	I can do the problems on math tests	−0.10 (0.16)	0.10 (0.44)	−0.02 (−0.19)	_0.63* (0.68)_	0.06 (0.25)
4	I can answer math questions that my teacher asks	0.03 (0.24)	0.20 (0.54)	−0.14 (−0.38)	_0.44* (0.55)_	0.02 (0.25)
5	I do well in math at school	−0.06 (0.29)	_0.78* (0.78)_	0.05 (−0.29)	0.14 (0.52)	−0.04 (0.25)
6	I can finish my math homework by myself	0.07 (0.14)	0.08 (0.30)	0.00 (−0.22)	−0.00 (0.28)	0.17 (0.22)
7	I can finish my math class work by myself	−0.00 (0.10)	0.02 (0.36)	0.05 (−0.24)	0.11 (0.40)	0.04 (0.29)
8	In the future, I will probably get a good grade in math	−0.01 (0.26)	_0.38* (0.54)_	0.06 (−0.17)	0.01 (0.31)	0.04 (0.33)
9	In the future, I will probably be able to solve the problems on my math tests	0.04 (0.19)	−0.03 (0.29)	−0.00 (−0.12)	0.13 (0.29)	0.20 (0.37)
10	In the future, I will probably be able to solve any new math problems I see	0.01 (0.18)	−0.07 (0.24)	0.06 (−0.07)	0.02 (0.20)	0.04 (−0.02)
11	When I get stuck with a math problem, I try new ways	0.11 (0.24)	−0.17 (0.00)	0.02 (0.08)	0.03 (0.07)	−0.10 (−0.02)
12	If a math problem is hard, I know where to get help	−0.01 (0.17)	−0.08 (0.07)	−0.01 (0.07)	−0.07 (0.06)	_0.41* (0.48)_
13	When I can't finish math by myself, I ask my teacher for help	−0.07 (0.13)	0.19 (0.12)	−0.05 (0.01)	−0.02 (0.04)	0.04 (0.14)
14	When I can't finish math by myself, I ask others (like my classmates) for help	0.10 (0.14)	−0.17 (−0.10)	0.00 (0.16)	0.12 (0.04)	0.01 (0.08)
15	If a math problem is new, I look for help in my book	0.01 (0.23)	0.03 (−0.06)	0.15 (0.23)	−0.01 (0.02)	−0.01 (0.05)
16	I check my answers before I turn in my math work	−0.02 (0.23)	0.06 (0.20)	0.03 (0.01)	0.15 (0.21)	0.08 (0.16)
17	I redo my math work if I started it the wrong way	−0.14 (0.06)	0.07 (0.18)	0.01 (−0.07)	−0.04 (0.10)	0.21 (0.26)
18	I reread the problem to understand it better	−0.13 (0.05)	0.09 (0.13)	−0.09 (−0.09)	−0.06 (0.05)	−0.07 (0.09)
19	I try to solve new math problems myself	0.10 (0.19)	−0.11 (0.20)	−0.08 (−0.13)	−0.02 (0.17)	_0.37* (0.41)_
20	I try to finish my math homework by myself	0.10 (0.19)	0.09 (0.23)	0.10 (−0.00)	−0.00 (0.21)	_0.49* (0.52)_
21	I try to finish my math class work by myself	−0.14 (0.03)	−0.02 (0.29)	−0.07 (−0.22)	−0.01 (0.25)	_0.58* (0.62)_
Mathematics Self-Concept (M-SC) items						
22	Is doing math class work fun for you?	_0.60* (0.66)_	0.09 (0.35)	0.07 (0.07)	0.05 (0.27)	0.12 (0.30)
23	Do you have fun with number work?	_0.50* (0.63)_	0.05 (0.30)	−0.01 (0.03)	−0.04 (0.19)	0.03 (0.21)
24	Do you enjoy playing math games?	0.26 (0.41)	−0.04 (0.17)	−0.07 (−0.00)	−0.07 (0.10)	0.04 (0.19)
25	Are you good at math?	0.03 (0.41)	_0.88* (0.86)_	−0.01 (−0.33)	0.04 (0.47)	−0.04 (0.27)
26	Have you always done well in math?	0.06 (0.28)	_0.60* (0.62)_	0.02 (−0.27)	−0.13 (0.25)	0.08 (0.25)
27	Do you think you get good grades in math?	0.02 (0.35)	_0.83* (0.82)_	−0.06 (−0.37)	0.02 (0.43)	−0.01 (0.28)
28	Do you think you are just as good in math as your classmates?	0.01 (0.28)	_0.62* (0.67)_	0.07 (−0.28)	−0.09 (0.35)	−0.09 (0.16)

(*continued*)

	Construct and items	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
29	Do you like to do math problems?	_0.74* (0.78)_	0.05 (0.37)	−0.08 (−0.03)	0.07 (0.28)	−0.04 (0.15)
30	Do you like to do math homework?	_0.79* (0.78)_	−0.06 (0.29)	0.02 (0.06)	0.02 (0.23)	0.03 (0.18)
31	Do you wish that math class was longer?	_0.69* (0.65)_	−0.05 (0.23)	0.07 (0.06)	0.02 (0.19)	−0.06 (0.05)
Mathematics Anxiety (M-ANX) items						
32	Do you feel afraid before math class?	−0.07 (0.00)	0.14 (−0.19)	_0.76* (0.66)_	−0.03 (−0.07)	−0.06 (−0.09)
33	Do you feel good before you take a math test?	_0.30* (0.37)_	0.16 (0.43)	−0.20 (−0.33)	−0.03 (0.23)	−0.05 (0.11)
34	Are you scared of math tests?	0.05 (0.09)	−0.10 (−0.28)	_0.63* (0.65)_	0.08 (−0.06)	−0.02 (−0.07)
35	Do you feel excited before math class?	_0.71* (0.72)_	0.08 (0.33)	−0.00 (−0.03)	−0.05 (0.17)	−0.04 (0.05)
36	Are you scared when the teacher asks you a math question?	0.08 (0.05)	−0.06 (−0.34)	_0.66* (0.70)_	−0.06 (−0.19)	0.02 (−0.08)
37	Do you want your teacher to ask you more math questions?	_0.59* (0.59)_	−0.00 (0.31)	−0.02 (−0.08)	−0.07 (0.19)	0.01 (0.15)

Notes: EFA on Year 1 Data Set 1 ($N_{Yr1\text{-}1}$ = 255). Loadings >0.30 in italics; Item 4 loading was 0.24 in one sample but 0.42 or above in cross-validation samples; hence, the item was retained. Promax rotated, sorted principal axis factor pattern shown with pattern coefficients and structure coefficients (in parenthesis)

Table AI.

Appendix 2

Characteristics of the students in the analytic samples

	Total sample (N_{Yr1} = 506)		Data Set 1 ($N_{Yr1\text{-}1}$ = 255)		Data Set 2 ($N_{Yr1\text{-}2}$ = 251)	
Variable	Frequencies	(%)	Frequencies	(%)	Frequencies	(%)
Female	270	54.2	134	53.2	136	55.3
Grade level						
5th Grade	269	53.2	130	51.0	139	55.4
6th Grade	237	46.8	125	49.0	112	44.6
Ethnicity group						
White	68	13.6	28	11.2	40	16.0
Black	292	58.5	146	58.6	146	58.4
Hispanic	91	18.2	52	20.9	39	15.6
Asian	46	9.2	23	9.2	23	9.2
Native America	2	0.4	0	0	2	0.8
Free/reduced lunch	243	62	127	64.5	116	58.9

Table AII.
Sample composition for Year 1 (N_{Yr1} = 506)

Table AIII.
Sample composition for year 2 *($N_{Yr2} = 469$)*

Variable	Total sample (N_{Yr1} = 469)		Data Set 1 ($N_{Yr1\text{-}1}$ = 234)		Data Set 2 ($N_{Yr1\text{-}2}$ = 235)	
	Frequencies	(%)	Frequencies	(%)	Frequencies	(%)
Female	256	54.6	126	53.8	119	50.6
Grade level						
5th Grade	215	45.8	107	45.7	118	50
6th Grade	254	54.2	127	54.3	117	50
Age level[a]						
11	134	28.6	65	28.3	69	28.9
12	254	54.2	128	55.6	125	52.8
13	68	14.5	38	17.0	38	15.8
14	7	1.5	3	1.7	3	1.3
Ethnicity group						
White	38	8.1	21	9.4	17	9.1
Black	298	63.5	151	63.7	147	61.8
Hispanic	90	19.2	39	17.5	51	21.4
Asian	40	8.5	20	8.5	20	7.5
Native America	3	0.8	3	0.9	0	0
Free/reduced lunch	291	62	141	60.3	141	60.2

Notes: [a] Age level represents student's age as of 09/01/2007; participation in free/reduced lunch school program is a binary indicator of poverty levels of student families

Corresponding author
Madhabi Chatterji can be contacted at: mb1434@tc.columbia.edu