ETS

Measuring the Power of Learning.™

Research Report
ETS RR–16-19

Development of the Mission Skills Assessment and Evidence of Its Reliability and Internal Structure

Kevin T. Petway II

Samuel H. Rikoon

Meghan W. Brenneman

Jeremy Burrus

Richard D. Roberts

May 2016

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Development of the Mission Skills Assessment and Evidence of Its Reliability and Internal Structure

Kevin T. Petway, II[1] Samuel H. Rikoon,[1] Meghan W. Brenneman,[1] Jeremy Burrus,[2] & Richard D. Roberts[2]

1 Educational Testing Service, Princeton, NJ
2 Professional Examination Services, New York, NY

The Mission Skills Assessment (MSA) is an online assessment that targets 6 noncognitive constructs: creativity, curiosity, ethics, resilience, teamwork, and time management. Each construct is measured by means of a student self-report scale, a student alternative scale (e.g., situational judgment test), and a teacher report scale. Use of the MSA provides schools with the opportunity to examine and monitor development of noncognitive skills in their students from Grade 6 to Grade 8. The use of the MSA has facilitated changes to curricula and more targeted instruction. The MSA scales exhibited meaningful relationships with standardized test scores and absenteeism, suggesting that these constructs play an important role in student behaviors and academic performance. This report presents reliability and factor analysis evidence for the student self-report and teacher report scales of the MSA to highlight its ability to measure the 6 constructs.

Keywords Noncognitive constructs; teacher report; student development

doi:10.1002/ets2.12107

The Mission Skills Assessment (MSA) is the result of a collaboration between the Educational Testing Service (ETS) and the Independent School Data Exchange (INDEX; formerly the Elementary School Research Collaborative). INDEX is a not-for-profit organization that facilitates the exchange of educational information among member schools serving students from kindergarten through either 8th or 12th grade. It is currently affiliated with more than 100 independent schools across the United States.

One common element among INDEX schools is their mission to establish an effective educational system fostering not only knowledge and skills but also the development of student character traits. To INDEX, a successful school is one that (a) increases student knowledge (e.g., facts, concepts, information), (b) improves academic and technical skills, and (c) develops positive character traits. The MSA was created to help address the latter issue of character, or noncognitive, development. *Noncognitive constructs* is a blanket term encompassing those attitudes, traits, behaviors, competencies, and skills (e.g., personality, communication, creativity) not traditionally measured by cognitive tests (e.g., standardized achievement tests). The research literature suggests that noncognitive constructs are more pliable than once believed (Durlak et al., 2011) and may actually change more readily than cognitive constructs (Kyllonen, Roberts, & Stankov, 2008). For example, Roberts, Walton, and Viechtbauer (2006) concluded, via a meta-analysis, that, on average, personal qualities, such as social vitality, social dominance, agreeableness, conscientiousness, emotional stability, and openness to experience, tend to change throughout the life span. Similarly, interventions targeting specific attitudes and skills, such as test anxiety, communication, resilience, and self-efficacy, can positively affect the expressed levels of these constructs in student samples (e.g., Greenberg et al., 2003; Hembree, 1988; Matthews, Zeidner, & Roberts, 2012).

Evidence also suggests that noncognitive constructs are predictive of academic achievement and future success. Wang, MacCann, Zhuang, Liu, and Roberts (2009) found meaningful relationships between teamwork skills and grades in both science and math for high school students. Similarly, Cohen (2006), as well as Ross, Powell, and Elias (2002), tied noncognitive constructs to successful interpersonal and professional development, decision making, and well-being. Outside of school, noncognitive constructs have predicted job performance (Barrick, Mount, & Judge, 2001), happiness (Diener & Lucas, 1999), health behaviors (Ajzen, Albarracin, & Hornik, 2007; Bogg & Roberts, 2004), marital satisfaction (Watson, Hubbard, & Wiese, 2000), and peer relationships (Jensen-Campbell et al., 2002), among others.

Corresponding author: K. T. Petway II, E-mail: kpetway@ets.org

The purpose of the current report is to provide a thorough summary of the MSA's development. To start, we describe the processes used to select the core constructs, to identify and develop the items to measure them, and to determine what types of additional data to collect. Following, we present results of a series of psychometric studies, including reliability analyses and multilevel confirmatory factor analyses (ML-CFA). We then discuss relationships between MSA constructs and both student demographic characteristics and standardized test scores. The report concludes with an evaluation of the current limitations of the MSA and proposed future avenues for research.

Overview

Core Construct Selection and Measurement

ETS staff relied on two key questions to guide selection of the core MSA constructs:

- What do staff at INDEX schools most want to know about their students?
- Which constructs are expected to relate to important student outcomes?

The name, Mission Skills Assessment, came about because of an initial survey of mission statements from member schools. Many INDEX schools, via their missions, claimed to value and develop a common set of noncognitive constructs in students: curiosity (i.e., a love of learning), work ethic, integrity, collaboration, and creativity. ETS also referenced influential reports such as *Are They Really Ready to Work?* (Casner-Lotto & Barrington, 2006), which surveyed employers to determine what applied and knowledge skills were necessary to succeed in the workforce and life in general. The report identified the following noncognitive constructs as those employers considered important for future success in the workforce: critical thinking, teamwork, creativity, work ethic, social responsibility, and curiosity, among others. These dominated the rankings made by employers, while traditional knowledge areas like history, English language, and even mathematics ranked quite a bit lower. Five of the six MSA constructs were selected based on the aforementioned sources: creativity, curiosity, ethics, teamwork, and time management (which incorporates elements of work ethic). Resilience was also included because research suggested constructs like or strongly related to resiliency (e.g., locus of control, self-efficacy, and grit) were important predictors of outcomes such as college completion, employment opportunities, employee pay, and general well-being (e.g., Duckworth, Peterson, Matthews, & Kelly, 2007; Heckman, Stixrud, & Urzua, 2006; Morrison, Brown, D'Incau, O'Farrell, & Furlong, 2006). Operational definitions that were used to guide the selection and development of items intended to measure each construct are provided in Table 1.

Apart from construct definitions, the creation and identification of scales used to measure each of the preceding constructs were motivated by Campbell and Fiske's (1959) multitrait–multimethod approach to evaluating construct validity. The multiple trait component is covered by the inclusion of the six theoretically related but conceptually different core constructs. Three methods of measurement were incorporated into the MSA for each construct: traditional Likert-type self-report items, Likert-type teacher report items, and an alternate type of assessment (e.g., situational judgments). Although all three sources of information are valuable, the current report focuses only on the self-report and teacher report components. As a result, this report describes a multitrait–multirater approach to construct validity, because self-report and teacher report utilize the same method (i.e., Likert ratings) but different raters. Additional research needs to be done with the alternate assessments, which are not discussed further in this report.

Assessment Methods

Likert-type self-report items are the most prevalent type of item in noncognitive assessment and make up the largest set of MSA items. Their key benefit is simplicity: They are straightforward to develop and test, and students generally respond to such items without difficulty. An example Likert-type item might ask a student to respond to a statement (e.g., "I like to play games") on a 5-point scale ranging from 1 (*not at all*) to 5 (*very much*). This type of item is susceptible to two forms of misrepresentation: reference bias and faking. Reference bias occurs when students rely on different pieces of information to make evaluations of themselves (e.g., King & Wand, 2007). Two students in different schools may have the same underlying trait level, but one may report lower levels than the other because their sources for comparison (e.g., other students in their schools) differ to some meaningful degree. This can lead to mean trait levels that differ due to response style rather than true mean differences. Reference bias is generally considered unintentional misrepresentation. In contrast, faking would be more akin to an act of intentional misrepresentation. Faking occurs when students deliberately choose a particular

Table 1 Operational Definitions for Each of the Six Core Mission Skills Assessment Constructs

Construct	Operational definition
Creativity	A student's ability to generate nonobvious ideas or solutions and willingness to engage with difficult or challenging problems (Carroll, 1993)
Curiosity	Self-motivated enthusiasm for learning and a desire to expand one's knowledge or skill set; often called "love of learning," which implies learning for its own sake (e.g., Berlyne, 1954; Nolen, 1988)
Ethics	A student's tendency toward trustworthy, dependable behavior. A student considered highly ethical is perceived as more altruistic, helpful, honest, and sensible (e.g., Moran, 2014).
Resilience	The capacity to function effectively despite encountered difficulties or stressors (MacCann, Duckworth, & Roberts, 2009). Resilience here also addresses a student's belief that he or she can recover or rebound from disappointments or failures (Judge, Erez, Bono, & Thoresen, 2003).
Teamwork	A student's ability to cooperate with others, influence others through support and encouragement, resolve conflicts or disagreements among group members, and guide or mentor other group members (Zhuang, MacCann, Wang, Liu, & Roberts, 2008)
Time management	Conscious control over the amount of time allotted to specific activities, usually through planning, to increase efficiency or reduce stress (Macan, Shahani, Dipboye, & Phillips, 1990)

response to present themselves as higher (or lower) on the trait than they actually perceive themselves to be (Ziegler, MacCann, & Roberts, 2012). This is more often a problem in high-stakes settings, where a student's responses could impact him or her significantly (e.g., if the results of an assessment were used to select students into a special program or generate some reward), or in low-stakes assessments, where the questions asked address behaviors that are perceived as very negative (e.g., drug use, cheating). We consider faking behavior to be uncommon in low-stakes settings such as those in which the MSA is administered, where no individual student reports are produced.

Teacher report is a type of "other report" where a teacher is asked to evaluate a student on the traits of interest. As part of the MSA, teachers are asked to rate each of their students on an abbreviated set of Likert-type items identical to those administered to students (though framed from the teacher's point of view). They are also used to evaluate preadolescent students because reading level and a greater tendency to utilize the extreme responses might impact the accuracy of student responses to self-report assessments (e.g., Marsh, 1986; Mellor & Moore, 2014). Teachers can provide a more accurate evaluation of a student's actual skill level than a student's own evaluation, and teacher reports can be more accurate predictors of student outcomes (e.g., Kenny, 1994). However, it is often difficult to standardize the level of experience a teacher has with his or her students. Similarly, teachers from different classes (e.g., math, history) may produce different profiles of the same student based on that student's interest in the subject matter or performance in that class.

Item Selection, Development, and Complexity

Many of the Likert items in the MSA were obtained from the International Personality Item Pool (IPIP; Goldberg et al., 2006). However, for most of the core constructs, items were also obtained from other sources. Select teamwork items were taken from the IPIP as well as from a study by Zhuang et al. (2008) that explored the development of a teamwork assessment for high school students. Refer to the cited study for validity evidence. Additional teamwork items were written to fully capture the construct as defined in Table 1. Creativity and resilience items were almost entirely obtained from the IPIP, with several variations of those items drafted to expand the item pool. Though several items in the curiosity and ethics scales can be found in the IPIP, the majority of the items for both scales were created by ETS research staff. The MSA steering committee, a panel of administrators from a subset of INDEX schools that were early adopters of the MSA, also assisted in the development of these items. Finally, all time management items were developed by ETS research staff. Refer to Appendix A for examples of self-report items from the MSA. Given how few items were taken from existing measures, there is little evidence of reliability and construct validity for the MSA scales prior to the current report.

When appropriate, the language of items included from the IPIP was adjusted to account for the anticipated reading levels of middle school students. The estimated readabilities and grade levels for each of the core scales are in Table 2. In general, Likert items were short, with an overall mean of 8.3 words per sentence, 1.4 syllables per word, and 4.0 characters per word.

Table 2 Readability Indexes for Student Likert Items

Construct	F – K Ease	F – K grade	C – L index	Characters per word	Syllables per word	Words per sentence
Creativity	84.0	3.4	7.6	4.0	1.4	7.3
Curiosity	72.9	4.9	9.5	4.3	1.5	7.2
Ethics	87.0	3.3	6.5	3.8	1.3	8.7
Resilience	89.3	2.8	6.1	3.7	1.3	7.9
Teamwork	82.7	3.9	7.3	3.9	1.4	8.9
Time management	92.3	2.7	7.2	3.9	1.2	9.2

Note. Higher ease scores are indicative of simpler text. A 0 on grade indexes is equivalent to kindergarten-level text. Higher grades are equivalent to the associated numerical grades in primary, intermediate, and secondary school (e.g., a grade of 12 suggests a 12th-grade reading level). F – K = Flesch – Kincaid; C – L = Coleman – Liau.

Table 3 Additional Student Data Provided by Schools

Variable	Description
Verbal reasoning	From the Educational Records Bureau's (http://erblearn.org/) Comprehensive Testing Program (CTP 4). This test focuses on analytical, categorical, and logical reasoning, with an emphasis on the ability to understand and infer information from text. It is available to students in Grades 3 – 11.
Quantitative reasoning	From the CTP 4. This test evaluates a student's ability to compare mathematical expressions and apply mathematical concepts to problems. It emphasizes a large number of mathematics-related reasoning skills. It is available to students in Grades 3 – 11.
Absences	Number of times the student has a recorded absence from school. Both semester and cumulative absences are collected.

The Flesch – Kincaid (F – K) Ease (Kincaid, Fishburne, Rogers, & Chissom, 1975) test was used to evaluate readability. F – K Ease scores can range from an undefined low to 120, with higher scores indicating easier readability. The simplest sentence (i.e., F – K Ease = 120) would contain two monosyllabic words (e.g., "I can"); a very long, complex sentence could garner an extremely negative score (e.g., −500). Easily understood text for an average 12-year-old (i.e., sixth-grade student) should obtain an F – K Ease score of approximately 80.0 or higher. The F – K grade level and Coleman – Liau (C – L) index (Coleman & Liau, 1975) tests were used to assess grade level. The F – K test utilizes syllable count, whereas the C – L index focuses on character count. It is useful to present both because their emphasis on different properties of a sentence could lead to different grade-level estimates.

The estimated F – K Ease score for the entire pool of student Likert items was 84.0, with an F – K grade of 3.6 and a C – L index of 7.5. Overall, these indexes suggest that readability is appropriate for this sample; although in some cases, the C – L index suggests a higher level of complexity than might be expected given that the MSA targets middle school students. This was most apparent with curiosity items, which had lower readability ease and a higher grade-level estimate than the other constructs.

Additional Student Data

Schools participating in the MSA are expected to provide additional data about their students. This information is used to better understand the relationship between MSA constructs and variables that both ETS research staff and INDEX member schools feel are important to study in relation to noncognitive constructs. These variables include covariates such as absences and test scores. Table 3 presents the variables used in the current report and provide a brief description of each.

Administration

INDEX schools administer the MSA near the end of the fall semester, generally between October and December. This provides students with enough time to solidify an impression of their classes and the academic year and gives teachers a sufficient amount of student exposure to ensure that their ratings of students are based on a broad survey of student behavior. The assessment is not designed for a particular time of year, though, allowing for alternative schedules or possibly multiple assessments per year. Currently only one form of the assessment is available, with items that are randomized during administration.

While INDEX has only focused on Grades 6–8, the assessment can be used at a variety of grade levels (given sufficient student reading ability). In addition, member schools currently administer the MSA annually to all students in these grades. This allows schools to track students across grades to provide a better understanding of student development over time. Despite having this longitudinal data, the present report uses data from only the 2013 data collection because it had the largest sample of students and teachers.

The MSA is a Web-based, computer-administered assessment accessible on all major operating systems (i.e., Windows, Mac, and Linux) and Internet browsers (i.e., Chrome, Firefox, Internet Explorer, Safari, and Opera). Students and teachers log in to the myMSA Web site[1] using individualized credentials, completing the assessment during either regularly scheduled periods (students) or at their leisure (teachers).

Although students can access the myMSA Web site from anywhere, completion of the assessment takes place at school under the supervision of a teacher or other school staff member. Students typically need 45–60 minutes to complete the student assessment. Some schools have chosen to split the assessment into two 30-minute sessions, whereas others ask students to complete it in one sitting. Teacher ratings are provided by the student's homeroom teacher, and only one teacher rates a student. The average number of students per teacher in the Fall 2013 sample is approximately 10. It takes between 4 and 5 minutes for a teacher to rate a single student, meaning the average teacher commits approximately 40–50 minutes to completing the MSA for their students.

Reporting

Schools do not receive individual student scores. MSA scores are aggregated across students within a school for each construct and by rater type (i.e., student self-report and teacher report), producing 12 scores (two per construct) total. The MSA reports provided to schools summarize this aggregated information only. INDEX and ETS staff judged this approach to be appropriate during initial development of the MSA to avoid comparisons between individual students (which could have inadvertently raised the stakes of the assessment). This aggregation also allowed INDEX schools to focus on one of the primary goals of INDEX as a larger institution, which was to facilitate discussions between member schools about best practices.

Data Analysis

Sample

The data used to provide evidence for the reliability and validity of the MSA came from the 2013 fall wave of data collection. The sample included $N = 13{,}158$ students[2] from 65 independent schools across the United States. The mean within-school sample size was 203 with a range from 68 to 368. Female students accounted for 48.4% of the sample, and the majority of students who reported a valid ethnicity identified as Caucasian/European American (70.3%). The next largest ethnic group responding, Asian Americans, comprised 10.8% of the sample, whereas African Americans were the third largest group, with 7.4% of the sample. Approximately 94% of students ($N = 12{,}333$) provided a response to the ethnicity variable, and approximately 10.9% of those students responded with "not sure." In 2013 and past assessment waves, the ethnicity/race item in the student background questionnaire included two conflicting options for students who identified as White: European American and Caucasian. Owing to possible student confusion, leadership at INDEX schools suspected the true proportion of White students in this sample was larger than reported here. Participation was roughly equal across all three grade levels. Appendix B has a complete breakdown of student demographics (including the number of "not sure" and missing responses for ethnicity). There were 1,293 teachers overall, and approximately 95% (1,243) of those teachers provided student ratings. As a result, teaches rated approximately 10 students on average. Missing data are discussed in more detail in the next section. Although it was possible to determine whether students did or did not complete the assessment, it was not possible to ascertain whether the assessment was not available to all students within a particular school given the information collected.

Table 4 Number of Items Targeting Each Mission Skills Assessment Construct

Construct	Student survey	Teacher survey
Creativity	22	6
Curiosity	28	7
Ethics	22	6
Resilience	22	8
Teamwork	25	6
Time management	26	7
Total	*145*	*40*

Table 5 Scale-Level Missing Data in Mission Skills Assessment Responses

Construct	Complete, *N* (%)	Incomplete, *N* (%)
Student survey		
Creativity	11,702 (93)	946 (7)
Curiosity	11,679 (92)	970 (8)
Ethics	11,727 (93)	922 (7)
Resilience	11,761 (93)	879 (7)
Teamwork	11,742 (93)	896 (7)
Time management	11,744 (93)	896 (7)
Teacher survey[a]		
Creativity	12,324 (100)	0 (0)
Curiosity	12,328 (100)	0 (0)
Ethics	12,300 (100)	52 (0)
Resilience	12,301 (100)	53 (0)
Teamwork	12,343 (100)	18 (0)
Time management	12,328 (100)	33 (0)

[a] *N* = number of students rated.

Method

Instrumentation

Table 4 presents a count of the number of items within each section of the assessment, inclusive of both the student and teacher surveys. The current study incorporated responses from students on 145 distinct items and from teachers on 40 items. All items were presented in a standard 4-point Likert format ranging from 1 (*never or rarely*) to 4 (*usually or always*).[3] In both the student and teacher surveys, item content was split approximately evenly across the six core constructs. Fewer items were included on the teacher assessment out of consideration for the time required for each teacher to complete the survey for multiple students.

Survey Nonresponse, Data Preparation, and Multilevel Structure

Table 5 provides a summary of the extent of nonresponse to each block of items in the MSA student and teacher surveys. Given the relatively small proportion of records missing any item response data on the student assessment (and near-complete lack of missing data on the teacher survey), all observations containing any valid response data were used in the statistical models reported. Response coding was left as originally delivered to study participants, with the exception of negatively worded items, which were reverse keyed to ensure consistent interpretations across the assessment and simplify factor analytic results.

Because students are clustered within teachers, observed variance in MSA item responses is naturally driven by differences between both students and teachers. When data are nested hierarchically, statistical models that account for the lack of independence among observations clustered within higher level units should be used (Raudenbush & Bryk, 2002). Although earlier analyses of an assessment serving as a precursor to the MSA could not make use of multilevel analysis at the teacher level due to missing data (Rikoon, 2013), the current sample realized a near 100% response rate among teachers.

Table 6 Mission Skills Assessment: Average Intraclass Correlation by Construct

Construct	Student survey, *M* (*SD*)	Teacher survey, *M* (*SD*)
Creativity	.02 (.01)	.23 (.09)
Curiosity	.04 (.02)	.24 (.02)
Ethics	.03 (.01)	.20 (.08)
Resilience	.02 (.01)	.20 (.03)
Teamwork	.02 (.01)	.19 (.03)
Time management	.05 (.02)	.16 (.06)

To determine the proportion of item response variance attributable to student clustering within teachers, intraclass correlations (ICCs) were calculated for each item. These coefficients estimated the proportion of total variation attributable to between-teacher differences, where the remainder was observed between children within teachers/classrooms. Table 6 presents the results of this analysis, with average ICCs reported for each group of items by construct.

We expected that the vast majority of differences in student response data would be driven by individual differences between students (as opposed to differences between their classroom contexts). First reviewing the student survey, it is apparent that only a nominal portion (5% or less on average) of total item response variance resided at the teacher level. In other words, our expectation was confirmed that differences between teachers played only a minor role in student responses to the MSA. Also as expected, the MSA teacher survey data evidenced a substantial proportion of item variance attributable to between-teacher differences. This ranged from 16% to 23% on average across core MSA constructs and was taken as an indication of the importance of using multilevel models to account for between-teacher variance in subsequent analyses of student-level noncognitive constructs.

Factor Structure

Each MSA item was designed to target a particular noncognitive construct. This targeting constituted a series of hypotheses as to which latent construct would account for the reliable variance in observed data associated with each item. ML-CFA was used to test a measurement model for each construct, specified according to the earlier hypotheses (D'haenens, Van Damme, & Onghena, 2010; Reise, Ventura, Nuechterlein, & Kim, 2005). These models provided an overall assessment of the fit of each construct's proposed dimensionality to the data collected from both students and teachers. Given that the intended use of the MSA is to report school-level scores on the six constructs of interest, our ML-CFA models were specified to be unidimensional in nature at both the student and teacher levels. This strategy had the advantage of providing estimates of the amount of item-level variance accounted for specifically at the between-student level (i.e., between-teacher variation was partialed out). It also provided estimates of the level of concordance between student and teacher ratings at the student level, which were distilled of both measurement error and between-teacher variance.

All ML-CFA models were fit using Mplus 7.3 (Muthén & Muthén, 2012). These models considered ordered categorical item data and were estimated using mean- and variance-adjusted weighted least squares (WLSMV) estimation following the recommendation of Flora and Curran (2004). Acceptable model fit was indicated by several criteria, including a comparative fit index (CFI) and Tucker–Lewis Index (TLI) $\geq .90$ and root mean squared error of approximation (RMSEA) $< .08$ (Kline, 2005, 2010). Close model fit was indicated by a CFI $\geq .95$ and RMSEA $< .06$ (Hu & Bentler, 1999).[4]

Although it is the most widely reported method of assessing factor reliability, the traditional estimation of a scale's reliability using Cronbach's alpha is based on Pearson product-moment correlations, which underestimate the magnitude of bivariate relationships when applied to ordinal data. Cronbach's alpha is unbiased given the condition of tau equivalence among indicator items (Brown, 2006; Raykov, 1997, 2004), which could not be assumed in the current case. These conditions make the traditionally computed alpha statistic a less than ideal indicator of a congeneric scale's reliability in practice, in particular where the scale is composed of variables containing Likert-type response data. For these reasons, the reliability of noncognitive factors in the MSA data was estimated using a version of Cronbach's alpha developed to accommodate Likert scale response data (Gadermann, Guhn, & Zumbo, 2012; Zumbo, Gadermann, & Zeisser, 2007). More specifically, this version of alpha uses polychoric correlations to estimate the reliability of construct-specific sets of unobserved continuous variables assumed to underlie observed student and teacher responses to each MSA item. Acceptable reliability was defined using the standard criterion (i.e., $\alpha_o \geq .70$).

Relationships With Student Demographics and Standardized Test Scores

After fitting an ML-CFA model for each construct and ensuring its adequate fit to the data, the same model was run inclusive of regressions specified such that each student-level latent variable was predicted by a vector of covariates including student gender, age, and days absent from school. Also included as predictors were Educational Records Bureau quantitative and verbal reasoning scores.

Scoring

At the time of this writing, Mplus did not have the capability to generate factor scores based on multilevel models estimated using WLSMV. As a result, unit-weighted factor scores were estimated for both student- and teacher-reported items via the average of all item responses within a given dimension. Similar scores were also generated using the weighted average of item responses, where factor loadings from the ML-CFA models were used as weights in the scoring equation. However, because the loading-weighted factor scores correlated $\geq .95$ with the unit-weighted factor scores across all six constructs and because unit-weighted scores were less susceptible to random features of the current sample, unit-weighted scores were adopted for reporting correlations between noncognitive dimensions. It is important to note here that no fine-grained distinctions between students are made on the basis of MSA factor scores because they are only ever reported at the aggregate school level.

Results

Several minor modifications were made to the item sets specified in Table 4, which were initially considered in their entirety by our ML-CFA models. Two items were removed from the ethics model because of convergence problems upon initial estimation. This was due to both items exhibiting near-zero teacher-level variance. Three items were also removed from both the resilience and teamwork models because they exhibited inadmissible parameter estimates (i.e., negative residual variances) upon initial estimation. Table 7 presents final ML-CFA model fit statistics for each noncognitive construct.

All of the models demonstrated acceptable fit to the supplied MSA item response data, with the curiosity and ethics models exhibiting close fit according to the criteria outlined earlier (see the “Method” section). These results were taken to indicate that it would be justifiable to estimate an overall factor score for each student on each of the six MSA constructs. Such scores would then be used to estimate the extent of relationships between MSA constructs (not directly estimable within an ML-CFA framework, given the total number of 185 items under consideration) as well as an aggregated school-level score on each construct for external reporting purposes.

The average amount of student-level variance accounted for in our ML-CFA models as well as the student-level correlation between student- and teacher-rated noncognitive constructs is presented in Table 8. Forty-one percent of student-level item response variance was accounted for on average across all constructs, with each construct explaining 30% or more of the same variance within its specific item group. Turning to the teacher-reported items, significantly higher proportions of student-level variance were explained by each noncognitive construct, suggesting a higher degree of average intercorrelation between the abbreviated set of teacher-reported items (within each construct) than was evidenced across the broader set of indicators reported by each student. This result makes logical sense, given the limited content scope of the teacher-report items in comparison to the more diverse block of items administered to students.

Student and teacher ratings evidenced moderate correlations at the between-student level. It is notable that these estimates were all greater in magnitude than analogous figures reported for the precursor assessment to the MSA (Rikoon, 2013). This finding suggests it is important to have distilled the data of both measurement error and teacher-level variance in the estimation of such relationships.

Relationships With Student Demographics and Standardized Test Scores

Table 9 presents standardized regression parameters from conditional ML-CFA models where each student-level latent variable was regressed on student demographic characteristics, absenteeism, and standardized test scores. These models were estimated using methods identical to those presented in Table 7 but applied to the subsample of 5,122 students with complete data on all independent variables.

Table 7 Multilevel Confirmatory Factor Analysis Model Fit

Construct	N_s	N_t	χ^2	*df*	CFI	RMSEA
Creativity	12,994	1,229	33,770	698	.943	.060
Curiosity	12,995	1,229	19,309	1,118	.967	.035
Ethics	12,995	1,229	14,396	596	.958	.042
Resilience	12,993	1,231	18,911	646	.928	.047
Teamwork	12,996	1,234	32,486	698	.930	.059
Time management	12,996	1,234	64,851	988	.942	.071

Note. All models were specified to be unidimensional at both the student and teacher levels and were estimated using robust weighted least squares. CFI = comparative fit index; N_s = number of students; N_t = number of teachers; RMSEA = root mean squared error of approximation.

Table 8 Average Student-Level Item Variance Explained and Student–Teacher Correlations by Mission Skills Assessment Construct

Construct	Student survey, *M* % (*SD*)	Teacher survey, *M* % (*SD*)	r_{s-t}
Creativity	46 (9)	86 (3)	.26
Curiosity	43 (17)	85 (4)	.33
Ethics	48 (16)	79 (9)	.28
Resilience	30 (12)	62 (21)	.24
Teamwork	35 (22)	80 (13)	.26
Time management	44 (13)	84 (15)	.43

Note. $s-t$ = student–teacher. $p < .001$ for all correlations.

Table 9 Standardized Regression Models of Predictors of Student-Level Noncognitive Construct Factors

	Independent variable				
Construct	Gender	Age	Absences	Quantitative	Verbal
Student survey					
Creativity	.03* (.01)	−.10 (.10)	.03 (.01)	.05** (.02)	.17*** (.02)
Curiosity	.08*** (.02)	−.03 (.08)	−.01 (.01)	.07*** (.02)	.11*** (.02)
Ethics	.33*** (.01)	−.02 (.08)	.00 (.02)	.02 (.02)	.03 (.02)
Resilience	−.09*** (.02)	−.02 (.07)	−.01 (.02)	.08*** (.02)	.04* (.02)
Teamwork	.14*** (.01)	−.09 (.05)	.00 (.01)	.04* (.02)	.03 (.02)
Time management	.32*** (.01)	−.01 (.03)	−.04 (.02)	.06*** (.02)	−.01 (.02)
Teacher survey					
Creativity	.12*** (.01)	−.01 (.27)	−.03* (.01)	.16*** (.02)	.22*** (.02)
Curiosity	.21*** (.02)	.01 (.12)	−.07*** (.01)	.17*** (.02)	.17*** (.02)
Ethics	.29*** (.02)	−.06 (.18)	−.06*** (.02)	.08*** (.02)	.03 (.02)
Resilience	.13*** (.02)	.00 (.27)	−.06*** (.01)	.17*** (.02)	.11*** (.02)
Teamwork	.28*** (.02)	.02 (.17)	−.05*** (.02)	.12*** (.02)	.04* (.02)
Time management	.33*** (.02)	−.11 (.23)	−.10*** (.01)	.18*** (.02)	.11*** (.02)

Note. Standard errors are in parentheses. $N = 5{,}122$ students (537 teachers).
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

First considering the student survey, females reported significantly higher levels of all noncognitive constructs than did males, with the exception of resilience (on which they were lower). Most pronounced were gender differences in ethics and time management, where effect sizes were greater than 0.3 *SD*. There were no differences in self-reported levels by either student age or days absent from school. In terms of test scores, higher levels of quantitative reasoning were associated with higher levels of all noncognitive constructs, with the exception of ethics. Results for verbal reasoning were more mixed, with students scoring higher in that area also exhibiting higher levels of creativity, curiosity, and resilience than their lower achieving peers.

Turning to the teacher survey, results look similar to the student survey in terms of gender and test scores. Females tended to rate more highly than males in all six noncognitive areas, as did students demonstrating higher levels of quantitative reasoning. One curious difference between teachers and students occurred with the relationship between gender

Table 10 Reliability and Correlations Among Student- and Teacher-Report Mission Skills Assessment Scores

	Creativity	Curiosity	Ethics	Resilience	Teamwork	Time management
Creativity	.97, .95	.78	.56	.37	.54	.37
Curiosity	.68	.97, .95	.62	.37	.51	.49
Ethics	.43	.65	.96, .94	.30	.63	.49
Resilience	.44	.54	.53	.92, .88	.43	.39
Teamwork	.48	.67	.78	.60	.96, .90	.51
Time management	.39	.63	.52	.54	.58	.96, .94

Note. Number of student ratings = 12,566. Number of teacher ratings = 12,300. Pairs of values along the diagonal represent reliability (α_o) estimates, where the first value is for the teacher-reported scale and the second value is for the student-reported scale. Displayed above the diagonal are correlations among the student-reported MSA scales. Below the diagonal are correlations among the teacher-reported scales. All $p < .001$.

Table 11 Correlations Between Student- and Teacher-Report Mission Skills Assessment Scores

	Student survey					
Teacher survey	Creativity	Curiosity	Integrity	Resilience	Teamwork	Time management
Creativity	**.22**	.20	.10	.10	.12	.13
Curiosity	.19	**.26**	.19	.17	.17	.27
Integrity	.05	.15	**.21**	.09	.14	.22
Resilience	.06	.10	.08	**.18**	.13	.15
Teamwork	.06	.14	.21	.13	**.20**	.25
Time management	.08	.15	.16	.17	.14	**.34**

Note. $N = 11{,}886$. All $p < .001$. Bolded values represent monotrait–heteromethod correlations.

and resilience, which is positive with teacher ratings (i.e., females rated higher than males) but negative with self-ratings (i.e., males rated themselves higher than females rated themselves).

Teacher ratings of student noncognitive constructs were also positively associated with student scores in verbal reasoning across the noncognitive spectrum, with the noted exception of ethics. There were no significant differences by student age in teacher ratings of their noncognitive constructs. Finally, higher levels of absenteeism among students were associated with lower levels of teacher ratings on all six MSA constructs. Rephrased, students who were absent more often than their peers tended to receive less positive noncognitive construct ratings from their teachers.

Relationships Between Noncognitive Constructs

Two types of statistics are presented in Table 10. Along the diagonal are estimates of ordinal alpha (Zumbo et al., 2007) for both student- and teacher-reported items making up each noncognitive score. These show high levels of reliability ($\alpha_o \geq .85$) across all constructs, with teacher-reported scales demonstrating slightly higher values than their student-reported counterparts. This finding may be due to the more homogeneous nature of item content within the abbreviated teacher scales.

Displayed above the diagonal in Table 10 are correlations among the student-reported MSA scores ($M = .49$, $SD = .13$); below the diagonal are correlations among the teacher-reported scores ($M = .56$, $SD = .11$). Here we observe moderate correlations between all MSA constructs, with two exceptions: the strong ($r = .78$) correlation between teamwork and ethics within the teacher-reported data and the equally strong association between creativity and curiosity within the student-reported data.

Table 11 displays relationships between student-report and teacher-report MSA scales. Bolded values along the diagonal are monotrait–heteromethod correlations (Campbell & Fiske, 1959). These reveal the magnitude of the association between the same theoretical trait assessed using different methods (here different raters).

It is clear that we observe markedly weaker associations between data sources than were exhibited by the within-method correlations reported in Table 10, which should not come as a surprise in light of the extant literature demonstrating low to modest student–teacher agreement on ratings of personality traits, self-concept, and social skills (Baker, Victor, Chambers, & Halverson, 2004; Gresham, Elliott, Cook, Vance, & Kettler, 2010; Laidra, Allik, Harro, Merenäkk, & Harro, 2006; Marsh, Smith, & Barnes, 1983). Of note in Table 11 is that for three of the six constructs (creativity, resilience, and time management), the monotrait–heteromethod correlation is greater than any other relationship involving the focal construct. In the case of curiosity and ethics, the difference between the monotrait–heteromethod value and the only other larger association is inconsequential ($\Delta r = .01$).

Discussion

This report has presented a detailed overview of the development of the MSA. Through an evaluation of reliability, factor structure, and predictive relationships, it has established the MSA as a viable tool for the assessment of noncognitive constructs important to student success in school, the workforce, and life in general. Though a considerable amount of additional research should and will be conducted to provide a more complete picture of the MSA, to INDEX member schools, the assessment provides a useful way to characterize their students with regard to the measured constructs. Schools, communities, educators, researchers, and policy makers armed with such data will be more informed in their work to determine how best to motivate positive development in all students.

Limitations and Future Directions

Although the readability indexes suggested the Likert statements were grade-level appropriate, in some cases, even monosyllabic words can be difficult for students to comprehend. Statements incorporating these words could be considered more readable than they really are by the tests used here. Several think-aloud sessions were conducted with students during the initial phase of the MSA's development to identify problematic terminology and statements. Results from those sessions did not suggest there were issues for this population (i.e., middle school students from participating MSA schools); however, studies conducted with a similar population of students for a different project did suggest there were grade-level differences in the interpretation of some disyllabic words. More thorough think-aloud sessions could be conducted with this population to ensure the item content is interpreted properly across all grade levels.

Given potential differences in interpretation by grade level, it may be useful to investigate the structural stability of these constructs across grade levels. It is possible that misfit for one group (e.g., sixth graders) is masked by a relatively good fit for the other two groups when the groups are not separated. The data collections for 2013 and subsequent years are conducive to multiple-group CFAs because of their large sample sizes, so future study will evaluate group differences by grade as well as other grouping variables, such as gender (and potentially ethnicity).

One of the primary findings demonstrated by Tables 8 and 11 is the limited degree of convergent validity observed between teacher- and student-reported measures of noncognitive constructs. The question remains, however, as to why student and teacher ratings of noncognitive constructs exhibited relatively low levels of agreement. Data collected in the current study were not capable of shedding light on this question directly, but several potential explanations highlighted by other researchers are worth mentioning for their relevance to the adolescent context. A review of the literature by Laidra et al. (2006) hypothesized that teachers might be less accurate judges of personality traits with less obvious or direct connections to classroom behavior or academic performance. Laidra et al. also suggested that the social context of classrooms themselves may affect student behavior, causing it to differ substantially in comparison to how children act in other environments. This theory is echoed by Baker et al. (2004) and Gresham et al. (2010), the latter providing evidence of higher convergent validity where rater contexts were more similar (e.g., both in school vs. both at home). That finding makes intuitive sense and comports with the idea that observed differences may be due in part to teachers rating students relative to one another within the limited context of their classroom observations, whereas a student's self-rating is bound to reflect an amalgam of influences from the student's exposure to others and his or her interactions with peers, parents, teachers, other family, and so on (Marsh & Shavelson, 1985).

It is vital to note that the rigorous collection of data from multiple informants is important *regardless* of the observed levels of agreement between them (Eid & Diener, 2006). To the extent that rater variance is reliable (as shown in the current study), information from multiple informants and assessment methods provides a more comprehensive view of student noncognitive constructs compared with an impression based solely on student self-reports. Eid and Diener (2006) suggested several reasons observed convergent validity may fall short of expectations (e.g., different interpretations of scale items, varying observational contexts) and recommended that where high levels of convergent validity are desired, researchers should collect variables expected to account for method-specific variance. Unfortunately, a detailed study of environmental context could not be carried out in the current research because of a lack of sufficient data on the classroom setting and academic subject(s) under consideration during teacher ratings. Further research in this area will be important to the extent that the pattern of student–teacher agreement observed herein is replicated in future MSA data collections. Future research might also explore teacher–teacher agreement by evaluating the level of consistency between teachers of different subjects. One might expect higher correlations for certain constructs (e.g., time management) than for other constructs (e.g., creativity) because of the stability of a construct's expression across classroom subjects (e.g., math vs. English).

In addition to teacher-related congruence issues, future study would investigate the longitudinal stability of the MSA constructs. Rikoon (2013) conducted longitudinal measurement invariance tests and identified structurally stable scales across two time points. However, that study used data from the 2011 and 2012 collections, which had much smaller samples of students and schools than the more recent waves did. Given the size of more recent MSA data collections, it would be helpful to revisit these analyses to incorporate all waves of data collected thus far. Extending from this, additional research could include investigations of group differences in construct validity, group differences in longitudinal changes, and predictors of any observed changes within or across groups. As the sample of participating schools expands to include a more nationally representative sample of students, the analyses suggested here will become particularly important. A critical component of support for the importance of noncognitive constructs in K–12 education is the idea that these constructs are malleable (as or perhaps even more so than traditionally emphasized cognitive constructs). Longitudinal analyses (in particular, those evaluating one or more behavioral interventions) will permit researchers to address the issue of malleability, and a comparison of students across different ethnic, socioeconomic status, and school subgroups might reveal valuable information about specifically where change is more readily observable.

Currently the MSA includes an alternate assessment method for each of the six core constructs. These alternative assessments were not analyzed for this report; however, future research would examine their relationships with the self- and teacher reports of the six core constructs as well as their relationships with demographic characteristics and test scores. This report also does not discuss relationships with several other student variables that were requested from schools and students: (a) tardies, (b) financial aid status, (c) mathematics grade point average, (d) language arts grade point average, and (e) life satisfaction. Future study would explore the association between them and the MSA core constructs.

A prevailing limitation of the current project that affects the generalizability of the findings in this report is the nature of the current MSA sample. INDEX membership comprises entirely independent schools that predominately serve students from relatively affluent families. Although some INDEX schools admit students from financially stressed home environments, the number of such students accepted is generally small and varies considerably from school to school. A key goal moving forward is to pilot the MSA in public school systems serving low-income, underserved groups. This would enable examinations of the generalizability of our findings and the psychometric stability of MSA scales across demographic subgroups. Positive findings could significantly expand the applicability and practical utility of the MSA.

Acknowledgments

Segments of this report were adapted from the second author's doctoral dissertation.

Notes

[1] See https://missionsskillsassessment.org/.

[2] This represents the number of students with grade information. Some students did not provide useable response data and were not included in subsequent analyses. Later sections indicate the sample sizes for these analyses.

[3] Although the MSA includes sections containing situational judgments, biographical items, and fluency tasks, the analyses presented herein focus exclusively on the Likert item response data.

[4] It is noted that the 90% confidence interval for the RMSEA statistic is normally also considered in assessing overall model fit; however, that statistic is not reported by Mplus when estimating multilevel models using WLSMV.

References

Ajzen, I., Albarracin, D., & Hornik, R. C. (Eds.). (2007). *Prediction and change of health behavior: Applying the reasoned action approach*. Mahwah, NJ: Lawrence Erlbaum.

Baker, S. R., Victor, J. B., Chambers, A. L., & Halverson, C. F. (2004). Adolescent personality: A five-factor model construct validation. *Assessment, 11*, 303–315.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*(1–2), 9–30.

Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology*, *45*(3), 180–191.

Bogg, T., & Roberts, B. W. (2004). Conscientiousness and health-related behaviors: A meta-analysis of the leading behavioral contributors to mortality. *Psychological Bulletin, 130,* 887–919.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York, NY: Cambridge University Press.

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century US workforce*. Washington, DC: Partnership for 21st Century Skills. Retrieved from http://www.p21.org/storage/documents/FINAL_REPORT_PDF09-29-06.pdf

Cohen, J. (2006). Social, emotional, ethical, and academic education: Creating a climate for learning, participation in democracy, and well-being. *Harvard Educational Review, 76*, 201–237.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology, 60*, 283–284. doi:10.1037/h0076540

D'haenens, E., Van Damme, J., & Onghena, P. (2010). Multilevel exploratory factor analysis: Illustrating its surplus value in educational effectiveness research. *School Effectiveness and School Improvement, 21*, 209–235.

Diener, E., & Lucas, R. E. (1999). Personality and subjective well-being. In D. Kahneman, E. Diener, & N. Schwartz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 213–229). New York, NY: Russell Sage Foundation.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92,* 1087–1101.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82,* 405–432.

Eid, M., & Diener, E. (2006). Introduction: The need for multimethod measurement in psychology. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 3–8). Washington, DC: American Psychological Association.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466–491.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, and Evaluation, 17*(3), 1–13.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40,* 84–96.

Greenberg, M. T., Weissberg, R. P., O'Brien, M. U., Zins, J. E., Fredericks, L., Resnik, H., & Elias, M. J. (2003). Enhancing school-based prevention and youth development through coordinated social, emotional, and academic learning. *American Psychologist, 58,* 466–474.

Gresham, F. M., Elliott, S. N., Cook, C. R., Vance, M. J., & Kettler, R. (2010). Cross-informant agreement for ratings for social skill and problem behavior ratings: An investigation of the Social Skills Improvement System-Rating Scales. *Psychological Assessment, 22,* 157–166.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). *The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior* (Working Paper No. w12006). Retrieved from http://www.nber.org/papers/w12006

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47–77.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Jensen-Campbell, L. A., Rosselli, M., Workman, K. A., Santisi, M., Rios, J. D., & Bojan, D. (2002). Agreeableness, conscientiousness, and effortful control processes. *Journal of Research in Personality, 36,* 476–489.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2003). The core self-evaluations scale: Development of a measure. *Personnel Psychology, 56,* 303–331.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis.* New York, NY: Guilford Press.

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel* (Report No. RBR-8-75). Millington, TN: Naval Technical Training Command Research Branch.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, *15*, 46–66.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Kline, R. B. (2010). Promise and pitfalls of structural equation modeling in gifted research. In B. Thompson & R. F. Subotnik (Eds.), *Methodologies for conducting research on giftedness* (pp. 147–169). Washington, DC: American Psychological Association.

Kyllonen, P. C., Roberts, R. D., & Stankov, L. (Eds.). (2008). *Extending intelligence: Enhancement and new constructs*. New York, NY: Taylor and Francis.

Laidra, K., Allik, J., Harro, M., Merenäkk, L., & Harro, J. (2006). Agreement among adolescents, parents, and teachers on adolescent personality. *Assessment, 13,* 187–196.

Macan, T. F., Shahani, C., Dipboye, R. L., & Phillips, A. P. (1990). College students' time management: Correlations with academic performance and stress. *Journal of Educational Psychology, 82,* 760–768.

MacCann, C., Duckworth, A. L., & Roberts, R. D. (2009). Empirical identification of the major facets of conscientiousness. *Learning and Individual Differences, 19,* 451–458.

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology, 22*(1), 37–49.

Marsh, H. W., & Shavelson, R. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist, 20*(3), 107–123.

Marsh, H. W., Smith, I. D., & Barnes, J. (1983). Multitrait–multimethod analyses of the Self-Description Questionnaire: Student–teacher agreement on multidimensional ratings of student self-concept. *American Educational Research Journal, 20,* 333–357.

Matthews, G., Zeidner, M., & Roberts, R. D. (2012). Emotional intelligence: A promise unfulfilled? *Japanese Psychological Research, 54*, 105–127.

Mellor, D., & Moore, K. A. (2014). The use of Likert scales with children. *Journal of Pediatric Psychology, 39,* 369–379.

Moran, S. (2014). Introduction: The crossroads of creativity and ethics. In S. Moran, D. Cropley, & J. C. Kaufman (Eds.), *The ethics of creativity* (pp. 1–22). Houndsmill, England: Palgrave Macmillan.

Morrison, G. M., Brown, M., D'Incau, B., O'Farrell, S. L., & Furlong, M. J. (2006). Understanding resilience in educational trajectories: Implications for protective possibilities. *Psychology in the Schools, 43*(1), 19–31.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén and Muthén.

Nolen, S. B. (1988). Reasons for studying: Motivational orientations and study strategies. *Cognition and Instruction, 5,* 269–287.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21,* 173–184.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy, 35,* 299–331.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment, 84,* 126–136.

Rikoon, S. H. (2013). *Toward an omnibus assessment of noncognitive skills: A longitudinal multitrait–multimethod application in middle schools* (Unpublished doctoral dissertation). Retrieved from ProQuest Dissertations & Theses database. (UMI No. 3594846)

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132,* 1–25.

Ross, M. R., Powell, S. R., & Elias, M. (2002). New roles for school psychologists: Addressing the social and emotional learning needs of students. *School Psychology Review, 31,* 43–52.

Wang, L., MacCann, C., Zhuang, X., Liu, O. L., & Roberts, R. D. (2009). Assessing teamwork and collaboration in high school students: A multimethod approach. *Canadian Journal of School Psychology, 24,* 108–124.

Watson, D., Hubbard, B., & Wiese, D. (2000). General traits of personality and affectivity as predictors of satisfaction in intimate relationships: Evidence from self- and partner-ratings. *Journal of Personality, 68,* 413–449.

Zhuang, X., MacCann, C., Wang, L., Liu, L., & Roberts, R. D. (2008). *Development and validity evidence supporting a teamwork and collaboration assessment for high school students*. Unpublished manuscript.

Ziegler, M., MacCann, C., & Roberts, R. D. (Eds.). (2012). *New perspectives on faking in personality assessment*. Oxford, England: Cambridge University Press.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*(1), 21–29.

Appendix A: Example Self-Report Items

Construct	Self-report item
Creativity	I think outside the box. I generate novel ideas.
Curiosity	I have a positive attitude about learning. I look forward to learning new things in school.
Ethics	I admit when I'm wrong. I think it is important to help people.
Resilience	I remain calm when I have a lot of homework to do. I get discouraged when things go wrong.
Teamwork	I cooperate with other students. I like to work with people.
Time management	I finish my homework on time. I am organized with my schoolwork.

Appendix B: Demographics for MSA 2013 Sample

Variable	Frequency	Percentage
Gender		
Female	6,366	48.4
Male	6,707	51.0
Missing	80	0.6
Grade		
Sixth	4,197	32.0
Seventh	4,400	33.4
Eighth	4,556	34.6
Missing	0	0.0
Ethnicity		
African or African American	803	6.1
Latino or Hispanic	301	2.3
Asian or Asian American	1,176	8.9
Native American	166	1.3
Caucasian or European American	7,664	58.2
Pacific Islander	68	0.5
International	411	3.1
Multiracial	311	2.4
Not sure	1,433	10.9
Missing	740	5.6
Age (years)		
<11	194	1.5
11	2,640	20.1
12	4,096	31.1
13	4,285	32.6
14	1,480	11.2
>14	3	0.0
Missing	460	3.5

Note. $N = 13{,}158$.

Suggested citation:

Petway, K. T., II, Rikoon, S. H., Brenneman, M. W., Burrus, J., & Roberts, R. D. (2016). *Development of the Mission Skills Assessment and evidence of its reliability and structure* (ETS Research Report No. RR-16-19). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12107

Action Editor: John Sabatini

Reviewers: Brent Bridgemanand Heather Busick