ets

2025

Pilot Validity Report: Health Sciences Program

Table of Contents

Tables and Figures	3
Executive Summary	4
Introduction	4
Part 1: Assessment Background	5
Presentation Format	5
Content	6
Scoring	8
Reporting	8
Part 2: Scale Correlations and Reliability	8
Intracorrelations of PSQ Dimensions and Composites	9
Part 3: Prediction Analyses	9
Part 4. Subgroup Comparisons	9
Part 5: Student Reactions to the Assessment	10
Part 6: Summary	12
Summary of Study	12
Potential Use Cases	13
Your reactions to the local validity study	13
References	14
Appendix A: Scoring	16
Appendix B: Reporting	17
Appendix C: Correlation Matrix of PSQ Dimensions and Composites	18
Appendix D: Prediction Analyses	20
Can a Common Prediction across Graduate Schools be Assumed?	20
Which PSQ variables are important?	21
Appendix E: Subgroup Differences	23

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

Tables and Figures

Figure 1. Sample PSQ Item (triplets version)	6
Table 1. Description of PSQ Dimensions and Composites	7
Table 2. Number of Pilot Participants across Schools	8
Table 3. Results of Post-Assessment Survey	11
Table 4. A Sample of Comments from School 1 Students Regarding the PSQ Assessment	12
Table 5. Intercorrelations of PSQ Dimensions and Composites for the Pooled School Sample	18
Table 6. Correlations for PSQ Dimensions and Student Outcomes for the Pooled School Sample	19
Table 7. Regression Analyses of Graduate Grades to Determine Appropriateness of Pooling	21
Table 8. 10-fold Cross-Validation to Determine PSQ Variables Importance	22
Table 9. Subgroup Mean Difference Effect Sizes across Gender, Racial and Ethnic Groups	23

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

Executive Summary

ETS's Personal Skills and Qualities (PSQ) assessment measures 14 dimensions (personality traits) that have been shown in prior research to be associated with successful performance in law school, professional school, or higher education more generally. PSQ is designed to be less susceptible than other behavioral assessments to score inflation due to coaching and faking. The assessment takes about 20 minutes to complete. The 14 dimensions are combined into composite scores measuring self-regulation, engaging with others, teamwork, emotion regulation, and innovativeness.

In this pilot study, 68 Graduate Students from a Health Sciences Program (refered to as "School 1" hereafter) took the assessment along with an additional 638 students from 8 other graduate and professional schools. School 1 sample data were pooled with data from the other schools to provide more stable results than might be obtained from conducting separate analyses by school.

A statistical analysis found that PSQ scores added substantially to the prediction of cumulative grade-point average over and above the prediction given by undergraduate grade-point average and standardized test scores. Analyses of score differences between certain subgroups for which there were enough respondents to conduct the analysis (male-female; African American-White) showed that subgroup differences were mostly small and statistically nonsignificant. Students' responses to the assessment were mostly favorable, citing the role PSQ can play in identifying student qualities that otherwise might be overlooked, and also citing the fairness of the PSQ in reflecting student strengths. Students also suggested that PSQ results could aid with development such as assisting in writing their personal statements.

Introduction

This report summarizes findings from a pilot test of the **ETS Personal Skills and Qualities (PSQ)** assessment, which was administered to 68 School 1 students during the period June - July 2021. PSQ is an online assessment measuring soft skills (also known as interpersonal and intrapersonal or noncognitive skills) essential to graduate school and career success. PSQ is intended as an assessment for students applying to graduate school. PSQ was designed to be used as a selection tool and to withstand the validity challenges associated with "select-in" use, such as susceptibility to coaching and applicant faking. However, it can also be employed in other settings, such as for student development.

Prior research suggests that the skills measured in PSQ are related to student success in higher education as well as subsequent career success (Mammadov, 2022; Poropat, 2009; Zell & Lesick, 2022). In addition, expanding admissions practices to include the kinds of measures included in PSQ will likely promote equity and lead to more diversity in graduate school admissions (Kalsbeek et al., 2013; Sackett et al., 2001): The dimensions measured in PSQ tend to show smaller racial and ethnic subgroup differences than do standardized tests. Therefore, PSQ is ideally suited to be used in the context of a holistic admissions process that considered evidence for student readiness to succeed in graduate school.

This validity report comprises six parts. Part 1 describes the assessment and provides background on how items were developed, scored, and reported. Part 2 presents descriptive statistics on the measure. We report results from all schools that have participated in the pilot project as well as specific evidence for any differences between School 1 students and other students. Part 3 presents the results of a statistical analysis in which grades were predicted from PSQ scores, to determine the degree to which PSQ predicts outcomes. Here again, results from all

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

schools are reported with evidence for any specific deviations with School 1 students from the general findings. Part 4 presents the results of comparisons between gender and racial and ethnic groups; here data were pooled from many sources because there were insufficient numbers of subgroup members to report specific subgroup comparison findings from School 1. Part 5 summarizes student reactions to the assessment based on a post-assessment survey. Part 6 provides a summary of the study, outlines several potential use cases, and solicits your feedback on the study and report through a few questions.

Part 1: Assessment Background

The importance of soft skills (also known as interpersonal and intrapersonal skills) for success in graduate school, in higher education more generally, and in the workforce has long been recognized. Graduate school applicants are sometimes evaluated for these skills through subjective measures such as personal essays, letters of recommendation, and resumes. Occasional research studies have been conducted using rating-scale self-reports, but these are susceptible to coaching and faking to appear more favorable when administered in a ”select-in” setting. Multidimensional forced-choice, described next, has advantages over these other approaches in contexts such as graduate school admissions (Cao & Drasgow, 2019).

Presentation Format

ETS’s PSQ is a multidimensional forced-choice (MFC) assessment of soft skills. Thus, it is an alternative or supplemental assessment approach to essays, letters of recommendation, and resumes, as measures of soft skills. PSQ is designed to withstand the challenges of use in a ”select-in” setting, although it can also be used in developmental setting. In PSQ, the candidate is presented either two or three statements, as shown below, and is asked to choose the statement that “is MOST like you” and the one “that is LEAST like you.” (In the pairs version the candidate is presented two statements and asked to choose the one “MOST like you.”) Statements are equated for their “social desirability,” that is, the degree to which students are likely to believe that selecting the statement would give them an advantage in an admissions decision. Because statements are thus balanced, there is no obvious better or preferred answer, and it is consequently difficult to coach or game the assessment.

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

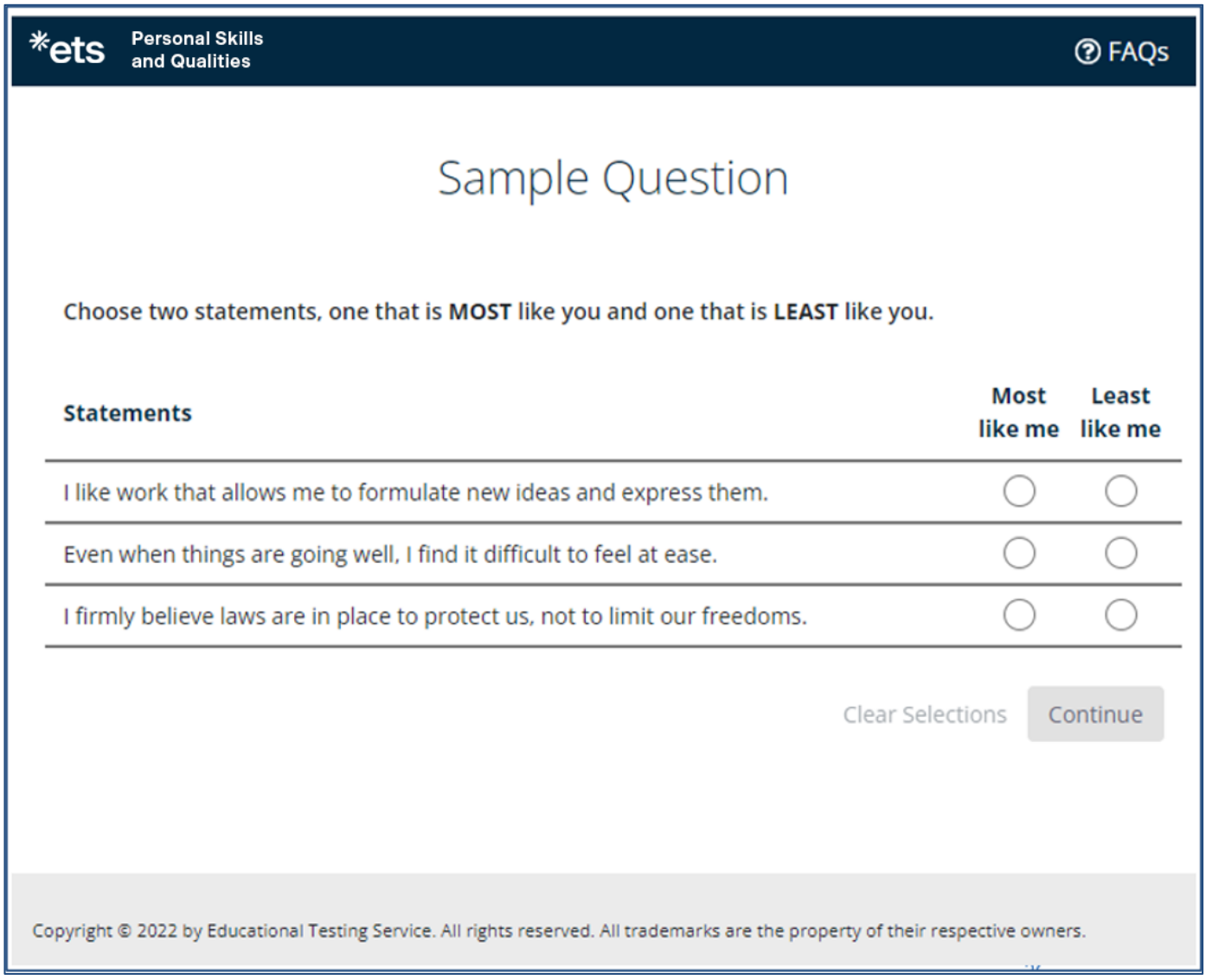


FIGURE 1. SAMPLE PSQ ITEM (TRIPLETS VERSION)

The triplets version of PSQ comprises 60 3-statement blocks (items); a comparable pairs version comprises 130 2-statement blocks. Recent research we have conducted has found that although the information extracted from the two forms of assessment is equivalent, the pairs format is more efficient and reliable and hence is the version that will be used going forward. The current version of the assessment is a fixed, linear form; research is underway to create an adaptive test version.

Content

The PSQ measures 14 dimensions, or scales, measuring attributes that have been identified in two ways: (a) through discussions with graduate school and other higher education and professional school faculty and staff members concerning the personal qualities most important for academic success, and (b) through a secondary analysis of hundreds of commercial scales of personal attributes, many of which have been identified as correlated with school outcomes as well as other higher education outcomes.

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

TABLE 1 IDENTIFIES THE 14 DIMENSIONS AND PROVIDES A DEFINITION OF EACH ONE. THE COMPOSITE DIMENSIONS ARE VARIOUS SUMS OF THE 14 DIMENSIONS (GOLDBERG, 1993).

TABLE 1. DESCRIPTION OF PSQ DIMENSIONS AND COMPOSITES

Dimension	**Description**	**Composite**
Achievement	Achievement refers to behaviors associated with diligence, ambition, hard work, goal striving, and proficiency.	Self-regulation
Organization	Organization refers to behaviors associated with punctuality, organization, and systematicity in work style.	Self-regulation
Responsibility	Responsibility refers to behaviors associated with loyalty, respecting obligations, and commitments, and being relied upon as a team member.	Self-regulation
Abidance	Abidance refers to behaviors associated with adherence to norms, traditions, and instructions from authority figures.	Self-regulation
Self-Discipline	Self-Discipline refers to maintaining focus, completing tasks, and considering options before deciding.	Self-regulation
Resiliency	Resiliency refers to internal psychological adjustment, a steady mood, and avoidance of worry even after negative feedback.	Emotion Regulation
Even-Temperedness	Even-Temperedness refers to being calm, level-headed, and good at regulating and navigating emotions even in stressful situations.	Emotion Regulation
Leadership	Leadership refers to behaviors associated with comfort in expressing opinions, leading, and being in charge in social contexts.	Engaging with Others
Sociability	Sociability refers to behaviors associated with comfort in approaching others and being interested in meeting new people.	Engaging with Others
Consideration	Consideration refers to behaviors associated with getting along with others and being a mediator or facilitator in group settings.	Teamwork
Trust	Trust refers to behaviors resulting from a belief in the goodness and trustworthiness of others.	Teamwork
Artistic Appreciation	Artistic Appreciation refers to behaviors associated with viewing and experiencing art and appreciating the beauty of and being inspired by art.	Innovativeness
Curiosity	Curiosity refers to behaviors associated with seeking out opportunities to learn and having varying interests.	Innovativeness
Creativity	Creativity refers to behaviors associated with coming up with new ideas and original solutions and enjoying engaging in unconventional thinking.	Innovativeness
Composite		
Self-regulation	Self-regulation describes the inclination to be diligent and responsible and to engage in behavioral self-control and self-management.	
Emotion Regulation	Emotion Regulation describes the ability to monitor and manage emotions and the behaviors associated with those emotions.	
Engaging with Others	Engaging with Others describes behaviors associated with how we socialize and influence other people.	
Teamwork	Teamwork describes behaviors associated with collaborating, cooperating, and building trust.	
Innovativeness	Innovativeness describes the desire to seek and use novel information and to produce or develop original methods and ideas.	

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

Scoring

Appendix A provides technical details on how the PSQ assessment is scored. We use a proprietary scoring algorithm which was shown in a comprehensive evaluation of possible scoring approaches we conducted to produce the most reliable scores and ones most likely to predict future outcomes.

Reporting

Composite scores as seen in Table 1 are created by summing their associated dimension scores. Descriptive statistics for composite and dimension scores are presented in the next section.
Dimension and composite scores are presented as five levels: high, medium+, medium, medium -, and low. Custom interpretation guides will be available for school personnel (i.e., those interpreting the scores for admissions or development uses) and for student/applicant assessment takers. The thresholds for the different score levels are determined by a norm group, specifically, all school applicants. The current PSQ system can produce such reports, and an evaluation of the usefulness of such reports is currently underway.

Part 2: Scale Correlations and Reliability

This section presents correlations among the 14 dimension scores and five composites and scale reliabilities. We report results from all schools participating in the pilot project as well as specific evidence for any differences between School 1 students and other students.

A total of nine schools participated in the pilot project resulting in a total pooled sample of 706 test takers. Table 2 lists the number of test takers participated in the pilot for each pilot school.

TABLE 2. NUMBER OF PILOT PARTICIPANTS ACROSS SCHOOLS

Pilot Schools	N
School 1	**68**
School 2	56
School 3	109
School 4	62
School 5	78
School 6	55
School 7	50
School 8	75
School 9	153

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

Intracorrelations of PSQ Dimensions and Composites

Appendix C shows the intercorrelations of the PSQ dimensions and composites based on the pooled school sample. As expected, the correlations of dimensions within corresponding composites (see Table 1) are relatively large, whereas the correlations of dimensions across composites are mostly small to moderate, supporting the proposed clustering (dimension into composite) structure listed in Table 1. The relationships are well-aligned with expectations. The intercorrelations based on the School 1-only sample are similar in pattern to those based on the pooled school sample. Sixty nine percent of the differences are within the range of -0.10 to 0.10, while the pooled sample generally has higher correlations between dimensions than the School 1-only sample

Part 3: Prediction Analyses

This section reports the results of analyses designed to determine the degree to which PSQ scores added to the prediction of student outcomes beyond the prediction given by the traditional predictors—undergraduate grade point average (UGPA) and standardized test scores, hereafter referred to as STS. To this point, the only student outcome we have available is graduate school grade point average (GSGPA), which was based on the most recent graduate school GPA for each participant in the study. School staff suggested that additional outcome variables were important such as faculty ratings, leadership positions and voluntary activities, and job offers upon completion, but these were not available at the time of the issuance of this report. Appendix D provides details of our analyses; here we provide only a brief summary.

For the analysis, first we determined that there was no differential prediction of PSQ across schools. That is, after adjusting for school differences, the same prediction model could be used for each of the pilot schools; we found no evidence that some predictors were more important at some schools than others. This is useful because this justifies pooling the data from the different schools to get a more accurate reading of the predictive power of the PSQ.

Next, we conducted an analysis to predict cumulative grade-point-average using STS scores and UGPA, and PSQ scores as predictors. We found that PSQ variables substantially improved the prediction of graduate school grades above and beyond the prediction given by the traditional predictors.

Part 4. Subgroup Comparisons

We conducted analyses of subgroup comparisons, for subgroups based on gender (female/male) and race and ethnicity (Black/White). There were insufficient numbers of participants from all subgroups to report specific findings from School 1, and insufficient numbers even after pooling all nine schools with missing demographic information from most schools. Therefore, to provide gender subgroup comparison data we pooled data from all the schools. To provide racial and ethnic subgroup comparison data, we pooled data from all these schools plus an independent sample of online workers and a sample of graduate and undergraduate students from a historically black college and university (HBCU) institution. We

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

realize that these are diverse samples, and even the selection of pilot schools that participated and returned student demographic information is likely not representative of the school population or the School 1 student population. Nevertheless, we believe there is value in these comparisons in that it provides some evidence for the degree to which the PSQ assessment shows differences between subgroups.

For **gender comparisons**, after pooling data from all nine schools, we obtained a sample of 222 female and 218 male students. Most of the effect sizes (differences between males and females) were smaller than 0.2 which is a small effect size (Cohen, 1988). Also, for some dimensions males were slightly higher; for others females were slightly higher. For six dimensions and one composite, the effect sizes were larger than 0.2, but the difference was only statistically significant for three dimensions and the composite that two of the dimensions contribute to, with the largest value being -0.49. These effect sizes are similar to those seen for large-scale standardized achievement tests in the areas of reading and writing (small effect sizes) and are smaller than those seen for achievement tests in the areas of math and science where male students on average outperform female students.

For **racial and ethnic comparisons**, after pooling as described above, we obtained a sample of 424 Back and 249 White students. As with gender, we found that most of the effect sizes were smaller than 0.20 indicating small effects across Black and White students. For most dimensions and composites, White students had slightly higher means than Black students, while for two of the dimensions, Black students had slightly higher means than White students. Consistent with expectations based on the literature, subgroup differences for PSQ were considerably smaller than subgroup differences on standardized tests.

Part 5: Student Reactions to the Assessment

After completing the PSQ assessment, students were given a post-survey questionnaire. The first part of the questionnaire asked students the extent to which they agreed with various statements concerning the importance, fairness, developmental uses, and usability of the PSQ.

Results are shown in Table 3.

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

TABLE 3. RESULTS OF POST-ASSESSMENT SURVEY

	% Agree/Strongly Agree	
	School 1 Students	All Students
Importance		
...schools should consider my skills and experiences beyond academic measures...when reviewing my graduate school application.	97	91
I feel that the kinds of skills measured in this assessment ...are important for success in a graduate program.	72	64
...were not captured elsewhere in my application to graduate school.	62	60
If admissions committees required an assessment like this one, it would give me a way to show skills and attributes that would be difficult for me to communicate elsewhere in my graduate school application.	54	49
Fairness		
The statements I was asked to evaluate ... seem equally fair to everyone.	62	64
...seemed unbiased.	62	64
Developmental Applications		
Getting feedback on my scores on this assessment could help me ...develop my interpersonal and intrapersonal skills.	71	80
...prepare for an admissions interview.	81	71
...write my personal statement.	72	69
Usability		
I liked being able to take this assessment on the device of my choice.	96	94
The assessment interface was simple to use.	91	88
This assessment was one of the quickest and easiest academic assessments I've ever taken.	43	40

The second part elicited open-ended responses to two questions:

- Describe the kind of feedback you would expect to get after taking this assessment.
- Describe what you liked best about this assessment.

Almost all students provided responses to these two questions. Responses varied and were generally favorable. Table 4 provides a sample of positive and negative responses to the second question. A complete list of the responses is available upon request.

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

TABLE 4. A SAMPLE OF COMMENTS FROM SCHOOL 1 STUDENTS REGARDING THE PSQ ASSESSMENT

"Describe what you liked best about this assessment"
Sample of positive comments
"Quick"
"I liked that it was simple and easy to use."
"The technology was easy"
"I liked that it asked a variety of questions on various topics"
"I liked that some of the choices were hard like if I was like hmm I don't think I'm any of these things or if I was struggling over thinking two of them matched me well but I could only choose one, because then I really had to reflect on what was most imporant to me."
"The assessment helped me think introspectively and challenged what i though about myself"
"I enjoyed the format"
"That kind of thought-exercising I find to be very enjoyable, and I hope that it captures an accurate reflection of not only who I think I am but who others think that I am as well."
"I liked that some of the questions challenged me and I had to actually stop and consider which of the three choices most and least described me (was especially difficult when all of them were applicable to me or none of them were)."
"I liked that the assessment addressed multiple aspects of my character"
"Easy-to-follow and clear format."
Sample of Negative comments
"I was frustrated that some sets of 3 statements all described me very much, and others I could not relate to a single one making the scoring very difficult"
"choose the best option is a good way to get someone but this wasn't necessarily easy"
"It felt somewhat limited and not a fair assessment of my beliefs/personality."
"It was simple, however, almost too simple to actually encompass my personality traits.."
"It was difficult having so many questions asked in multiple ways I felt like I was being tricked somehow.

Part 6: Summary

Summary of Study

The PSQ assessment is a novel assessment of 14 dimensions that have been shown in prior research to be associated with successful performance in higher education. The assessment was designed to mitigate score inflation due to coaching or faking. The assessment takes about 20 minutes to complete. PSQ's 14 dimensions can be combined into composite scores for reporting purposes, for example, as feedback to test-takers. We suggest a set of composites that reflects the natural clustering of the 14 dimensions into the five categories of self-regulation, emotion regulation, engaging with others, teamwork, and innovativeness. Other composites are possible as well, such as a composite designed to best predict student outcomes.

To evaluate the efficacy of the PSQ assessment we conducted prediction analyses to show that that PSQ scores added substantially to the prediction of graduate school grades above and beyond the prediction given by undergraduate grades and standardized test scores. Further analyses showed that gender and racial and ethnic subgroup differences were mostly small and nonsignificant. Students' responses to the assessment were mostly

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

favorable, citing the role PSQ can play in identifying student qualities that otherwise might be overlooked and the perceived fairness of the PSQ. Students also suggested that PSQ results could help them with developmental applications such as writing their personal statements.

Potential Use Cases

The PSQ assessment could be used in several ways by a school. It is important to distinguish high stakes use ("to provide results that have important, direct consequences for individuals…")[1] from low stakes use ("to provide results that have only minor or indirect consequences for individuals…")[2]. Medium-stakes use is commonly understood as falling between high- and low-stakes uses. Using PSQ as part of an admissions decision-making process would likely be characterized as a medium-stakes use, given that PSQ information would be considered alongside the traditional criteria of undergraduate grade-point average and standardized test scores. An early-in-the-process use would involve PSQ scores being considered for screening candidates. The advantage of using PSQ information for screening would be that such a use would simultaneously increase accurate identification of who would be most likely to succeed academically and identify candidates whose selection would increase the racial and ethnic diversity of the admitted class.

A late-in-the-process use would be to bring in PSQ information at the final stages of admissions decision-making for on-the-fence-case candidates—those who might or might not be admitted depending on information considered beyond the standard admissions measures. PSQ information could be used to decide between two candidates who otherwise seem similar in their merits for admissions.

For using PSQ this way, the 14 PSQ dimensions that the assessment provides could be combined into a single composite. This would not be done immediately, but instead would be done over time, after the accumulation of data from candidates and students and the experience of the admissions staff in understanding how scores could be used. There are several composites we would recommend potentially for operational use. An *academic composite* would be designed to reflect PSQ information in the form of a single score that would best predict likely academic success, as indicated by cumulative grade-point average. *A professional development composite* would be designed to predict likely success of the student on non-academic factors, such as contributions through leadership or voluntary activities or general contributions to the school community. A third, *balance composite* would reflect a combination of academic and nonacademic success and could be used in balancing the admitted class for both aspects. Again, these composites could be developed after a year or two of experience with the assessment.

An alternative to the use of PSQ information for admissions would be to use PSQ to provide developmental feedback to students or to monitor changes in class composition over the years with respect to PSQ scores. The former use might result in suggestions for activities and readings or could be a point of discussion in student and career counseling. The latter use might result in evaluations of effects of admissions policy changes, changing applicant demographics factors, and other trends that might affect the status of the program.

Your reactions to the local validity study

We would be interested in getting your reactions to the validity study. The following are questions for consideration:

1. Are the findings from the study clear? Are there aspects to the study you would like to hear more details about, or an elaboration on?
2. If you were to adopt PSQ for admissions, how would you use it?
3. If you were to adopt PSQ for a non-admissions application, such as for student development, how would you use it?

[1] AERA, APA, NCME (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association, p. 219.
[2] AERA, APA, NCME (2014). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association, p. 221.

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

References

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology, 104(11),* 1347–1368. https://doi.org/10.1037/apl0000414

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, R. A. (2006). Introducing the GLMSELECT PROCEDURE for Model Selection. SAS Institute, Inc. Available online at: http://www2.sas.com/proceedings/sugi31/207-31.pdf (accessed May 31, 2022).

Credé M, Kuncel NR. (2008). Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance. *Perspectives on Psychological Science, 3(6),* 425-453. doi:10.1111/j.1745-6924.2008.00089.x

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin, 68(3),* 161-182.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist. 48 (1),* 26–34. doi:10.1037/0003-066X.48.1.26.

Hurvich, C. M., & Tsai, C.-L. (1991). Bias of the corrected AIC criterion for underfitted regression and time-series models. *Biometrika, 78(3),* 499-509.

Kalsbeek, D., Sandlin, M., & Sedlacek, W. (2013). Employing noncognitive variables to improve admissions, and increase student diversity and retention. *Strategic Enrollment Management Quarterly, 1(2),* 132-150.

Kyllonen, P. C., Heincke, P., Holtzman, S., Olivera-Aguilar, M., Iocovelli, M., Xu, J., Grodman, L., & DelMonico, B. (2020). *Predicting Yale School of Management Outcomes with a Forced-Choice Behavioral Assessment.* Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Mammadov, S. (2022). Big Five personality traits and academic performance: A meta-analysis. *Journal of Personality, 90,* 222-255. https://doi.org/10.1111/jopy.12663

McElroy Tutoring (2007-2022). *GRE vs. GMAT vs. LSAT: The Ultimate Score Percentile Comparison / Conversion Table (Updated for 2019-20).* Retrieved from https://www.mcelroytutoring.com/blog-post.php?id=4887 on 10-31-2022.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322–338. https://doi.org/10.1037/a0014996

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138,* 353–387.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, *130*, 261-288.

Robbins, S. B., Allen, J., Casillas, A., Peterson, C. H., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology,98,* 598–616.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. American Psychologist, 56(4), 302–318. https://doi.org/10.1037/0003-066X.56.4.302

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

Sawa, T., (1978). Information criteria for discriminating among alternative regression models. *Econometrica, 46(6),* 1273-1291.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6 (2),* 461–464, doi:10.1214/aos/1176344136.

Zell, E., & Lesick, T. L. (2022). Big five personality traits and performance: A quantitative synthesis of 50+ meta-analyses. *Journal of Personality, 90*(4), 559–573. https://doi.org/10.1111/jopy.12683

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

Appendix A: Scoring

There are several ways responses to the PSQ assessment can be modeled to produce the 14 dimension-scores. *Simple sum scoring* entails a count of each time a statement reflecting a particular dimension is selected over the other statements(s) reflecting different dimensions. Sum scoring of pairwise comparison (or "forced-choice") response data yielded by the PSQ assessment does not allow comparisons between individuals, a condition known as *ipsativity*.[3] With ipsative data, it is possible to determine for each individual relative preferences or strengths and weaknesses on the various dimensions, but not to compare two individuals on those dimensions. But in recent years, several item-response theory (IRT) models have been proposed for scoring forced-choice response data that overcome the limitations of ipsative data and allow for comparisons between individuals.[4]

ETS conducted a comprehensive evaluation of these IRT approaches and added several new models and unpublished variations on these models to the evaluation. Models were evaluated by a variety of criteria including the internal consistency reliability of the various resultant scores, various model fit measures, and test-retest reliability assessments. The best model resulting from that evaluation is the model that was used to score the response data from School 1 and the basis for analyses presented in the main body of this report. The scoring model is now implemented in the scoring engine of the assessment. The results of the evaluation and description of the scoring model is documented and will be published.[5]
The referenced paper provides details.[7]

[3] Cattell, R. B. (1944). Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. Psychological Review, 51(5), 292–303. https://doi.org/10.1037/h0057299
[4] Stark, S., Chernyshenko, O.S., Drasgow, F. (2005). An IRT Approach to Constructing and Scoring Pairwise Preference Items Involving Stimuli on Different Dimensions: The Multi-Unidimensional Pairwise-Preference Model. *Applied Psychological Measurement, 29(3),* 184-203. doi:10.1177/0146621604273988; Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18,* 36-52. doi:10.1037/a0030641
[5] Fu, J., Tan, A., Kyllonen, P., Devasia, N., Holtzman, S. (2022). *A Comparison of Item Response Theory Models for Force Choice Questionnaires: Real Data Examples.* Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

Appendix B: Reporting

Response data modeling and scoring, described in Appendix A, results in 14 raw dimension scores for each respondent. The scale of each of the raw dimension scores is set to a mean of 0 and a standard deviation of 1, based on a large pool of test takers taking either of two forms of the PSQ under various circumstances. At this point the scale is not strictly interpretable with respect to a clearly identified norm group. At a later point these raw scores will be converted to a scale that will be more clearly identified with a specific norm group, such as applicants. Then, for example, a score of 0 will be interpretable as the mean score (for a dimension) of all applicants, and a score of -1 will be interpretable as a score one standard deviation below the mean of all applicants for that dimension.

Composite scores are created from the raw dimension scores by weighing them and summing the weighted scores. The composite scores listed in Table 1 are created as the weighted sum of the scale scores to which they are associated. A self-regulation score can be created by summing the unit-weighted first five dimension scores listed in TABLE. The other four Big 5 scores can be similarly created, based on the labels presented in the third column of TABLE. Results based on these Big 5 composite scores, along with the 14 dimension scores are presented in Part 2: Descriptive Statistics.

Composite scores can also be created based on a model that best predicts Business School Grade Point Average or other outcomes. Models that could serve as the basis for such a composite score are presented in the section Part 3: Regression Models.

With raw scores and with composite scores, score interpretation can be facilitated with the use of score categories, such as high, medium, and low (i.e., three levels) or high, medium+, medium, medium -, low (i.e., five levels). Then custom interpretation guides can be produced for either school personnel (i.e., those interpreting the scores for admissions or development uses) or for student/applicant assessment takers. The thresholds for the different score levels can be determined by a norm group, such as all student test-takers. The current PSQ system is capable of producing such reports, and an evaluation of the usefulness of such a report is currently underway.

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

Appendix C: Correlation Matrix of PSQ Dimensions and Composites

TABLE 5. INTERCORRELATIONS OF PSQ DIMENSIONS AND COMPOSITES FOR THE POOLED SCHOOL SAMPLE

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. Achievement	1.0																		
2. Organization	.61**	1.0																	
3. Responsibility	.77**	.68**	1.0																
4. Abidance	.39**	.42**	.42**	1.0															
5. Self-Discipline	.68**	.65**	.82**	.38**	1.0														
6. Resiliency	.30**	.12**	.26**	.06	.43**	1.0													
7. Even-temperedness	.18**	.08*	.25**	.05	.43**	.84**	1.0												
8. Leadership	**.59****	.25**	.32**	.05	.29**	.43**	.18**	1.0											
9. Sociability	.35**	.08*	.12**	.12**	.08*	.40**	.28**	.57**	1.0										
10. Consideration	.27**	.10**	.22**	.22**	.20**	.34**	.45**	.24**	**.62****	1.0									
11. Trust	.00	-.08*	.00	.10**	.07	.40**	**.51****	.09*	.44**	.66**	1.0								
12. Artistic Appreciation	-.06	-.06	-.12**	-.14**	-.07*	-.03	.05	-.02	.12**	.15**	.17**	1.0							
13. Curiosity	.17**	.03	.11**	-.20**	.21**	.21**	.27**	.20**	.14**	.20**	.20**	.51**	1.0						
14. Creativity	.28**	.03	.14**	-.19**	.20**	.34**	.25**	.45**	.32**	.22**	.19**	.48**	.63**	1.0					
15. Self-regulation	.87**	.84**	.92**	.46**	.89**	.31**	.26**	.42**	.18**	.22**	-.01	-.09*	.15**	.18**	1.0				
16. Engaging with Others	**.52****	.18**	.24**	.09*	.20**	.46**	.26**	.87**	.90**	**.50****	.31**	.06	.19**	.43**	.33**	1.0			
17. Teamwork	.15**	.01	.12**	.18**	.15**	.41**	.53**	.18**	**.58****	.91**	.91**	.17**	.22**	.22**	.12**	.44**	1.0		
18. Emotion Regulation	.25**	.11**	.27**	.06	.45**	.96**	.96**	.32**	.35**	.41**	.48**	.01	.25**	.31**	.30**	.38**	.49**	1.0	
19. Innovativeness	.16**	.00	.05	-.21**	.14**	.21**	.23**	.25**	.23**	.23**	.22**	.80**	.85**	.85**	.10*	.27**	.25**	.23**	1.0

*Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed). This table is based on all 9 schools, *N* = 706 who took PSQ.

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

TABLE 6. CORRELATIONS FOR PSQ DIMENSIONS AND STUDENT OUTCOMES FOR THE POOLED SCHOOL SAMPLE

Dimension/Composite	**GSGPA** (N=695)	**Admission Test Score Percentile** (N=418)	**UGPA** (N=366)
Achievement	0.16	-0.13	0.18
Organization	0.08	-0.14	0.19
Responsibility	0.16	-0.13	0.16
Abidance	0.07	-0.11	0.15
Self-Discipline	0.16	-0.14	0.18
Resiliency	-0.01	-0.03	-0.09
Even-temperedness	-0.01	0.01	-0.08
Leadership	0.07	-0.10	0.06
Sociability	-0.01	-0.11	-0.02
Consideration	0.01	-0.07	-0.03
Trust	-0.06	0.02	-0.13
Artistic Appreciation	-0.16	-0.03	-0.16
Curiosity	-0.10	-0.05	-0.19
Creativity	-0.15	-0.08	-0.17
Self-regulation	0.16	-0.15	0.20
Engaging with Others	0.03	-0.12	0.02
Teamwork	-0.03	-0.03	-0.09
Emotion Regulation	-0.01	-0.01	-0.09
Innovativeness	-0.16	-0.07	-0.20

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

Appendix D: Prediction Analyses

This section provides additional technical detail on the prediction analyses, the series of regression analyses designed to determine the degree to which PSQ scores added to the prediction of graduate school GPA (GSGPA) beyond the prediction given by the traditional predictors—undergraduate grade point average (UGPA) and standardized test scores (standardized test score percentiles based on GMAT, LSAT, and GRE, hereafter referred to as STS).

Three sets of regression analyses were conducted. The first addressed the issue of whether the graduate schools included in the sample could be pooled to stabalize regression coefficient estimates. The second analysis was designed to determine the importance of the specific PSQ dimension scores, and therefore the best way to assign a weight to the PSQ variables to make a best-prediction composite. This was done with a k-fold cross-validation approach. K-fold cross validation (the models presented here used 10-fold) involves fitting the model on a subset of the data (9/10) then applying the model parameters to the holdout set data (1/10) to determine model fit (computing error on the holdout set data), and repeating this step for each of 10 subsets. This way, all the data are used for the best prediction, but the danger of overfitting is reduced because data used for model fitting are not used to estimate the error in model fitting.
The third set of analyses used a gradient boosting machine (GBM) approach, a technique widely used in machine learning, which has proven to provide excellent prediction results in forecasting competitions.[6] For the application shown here, gradient boosting is applied to regression tree analysis. The prediction given by gradient boosted regression trees is often better than the ones given by the two other approaches, but the interpretation of that prediction is less transparent with GBM due to the complexity of the model. Studying the difference between the GBM prediction and the prediction given by the more transparent linear regression models may yield insights into how the prediction variables may be better utilized. Nevertheless, we found that GBM did not improve model fit on a hold-out sample and so we do not report the results of that analysis here.

For these analyses model fit is evaluated by *R*-squared, adjusted *R*-squared, and mean square error (MSE), and additional fit indices where appropriate.

Can a Common Prediction across Graduate Schools be Assumed?

The pooled graduate school dataset includes a total of nine graduate schools with 706 students. The School 1 dataset includes 68 students. Analyzing the pooled graduate school dataset will generally provide more stable (i.e., lower variance) estimates than the results from a single school. However, schools give grades on different scales and so we replaced grades with within-school percentiles from the grades provided. Next, we put standardized test scores (GRE, GMAT, LSAT) on a common scale using a published concordance table (McElroy Tutoring, 2007-2022). Thus we have a common outcome measure (grade percentiles) and a common standardized test score measure (concorded test scores) for the regression analyses.

Independently of the common measures issue, pooling the data can miss school-specific relationships, such as when the relationship between one of the PSQ dimensions and grades is higher in one school than it is in another (i.e., differential prediction). So, we first tested whether there was any evidence for school-specific relationships. Table 7 shows the results of the OLS regression designed to address this question. (From 9 schools and 706 students, three schools failed to report undergraduate grades and one of those three also failed to report standardized test scores, reducing sample size and the number of school effects examined.) The first model

[6] Putka, D. J., Beatty, A. S., Reeder, M. C. (2017). Modern prediction methods: New perspectives on a common problem. Organizational Research Methods, https://doi.org/10.1177/1094428117697041

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

column (Model 1) presents results when including only the school fixed effects, which adjusts for differences between schools in expected grades. The Model 2 column shows results of adding the two traditional admission predictors (UGPA, STS). The Model 2' column adds the school by traditional predictors to adjust for differential prediction across schools for the traditional predictors. The Model 3 column shows results after adding the 14 PSQ dimension scores to the prediction model. The Model 4 column shows results of including PSQ by school interaction variables to Model 3.

TABLE 7. REGRESSION ANALYSES OF GRADUATE GRADES TO DETERMINE APPROPRIATENESS OF POOLING

	N effects	Model 1	Model 2	Model 2'	Model 3	Model 4
Constant	1	1	1	1	1	1
School Fixed Effects (S)	5	5	5	5	5	5
Traditional Predictors (T)	2		2	2	2	2
S x T interactions	10			10		
PSQ Predictors	14				14	14
School x PSQ interactions	70					70
Model *df*		5	7	17	21	91
Residual *df*		371	340	330	326	256
MSE		846.4	644.5	639.8	576.4	583.5
SSE		31401	21914	21112	18790	14937
		9.8	4.8	1.1	4.3	8.1
F (model comparison)			73.60*	1.25	3.87*	0.94
R^2		0.00	0.23	0.26	0.34	0.48
adj R^2		-0.01	0.22	0.22	0.30	0.29

Notes. Outcome variable is graduate school GPA. *F* (model comparison) is the *F* test of the difference between each model and the model that precedes it (i.e., model 3 vs. model 2; model 2 vs. model 1). The other statistics are for the model and not comparisons. *F* tests marked with an "*" are significant at the 0.05 level.

F tests are computed for comparing adjacent models (I.e., Model 2 vs. Model 1; Model 2' vs. Model 2; Model 3 vs. Model 2'; and Model 4 vs. Model 3). It can be seen that (a) the F test for Model 2' (vs. Model 2) is not significant, suggesting there is no evidence for differential predication of STS and UGPA across schools. The *F* test comparing Model 4 with Model 3 is also not significant, indicating no effect for adding the school by PSQ interaction terms, and therefore it is reasonable to assume that there is no differential prediction of PSQ across schools. These findings justify the use of the pooled regression results for further exploration.
Also, comparing model 3 with model 2' (or 2) shows that there is a significant effect for the PSQ variables, suggesting that PSQ variables improve the prediction of graduate school grades above and beyond the prediction given by the traditional predictors.

Which PSQ variables are important?

This analysis used the same set of predictors as the previous OLS models (Models 2 and 3), but in addition polynomial and interaction terms for the 14 PSQ variables were added to capture nonlinear relationships between PSQ predictor variables and grades. Table 8 presents the findings and comparisons between models.

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

TABLE 8. 10-FOLD CROSS-VALIDATION TO DETERMINE PSQ VARIABLES IMPORTANCE

	N effects	Model 5	Model 6	Model 7
Constant	1	1	1	1
Traditional Predictors	2	2	2	2
School Fixed Effects	5	5	5	5
PSQ Predictors	14		5	5
PSQ Predictors (polynomials)	14		0	1
PSQ Predictors (PSQ interactions)	91		0	0
Model *df*	126	7	12	13
Residual *df*		340	335	334
MSE		644.5	569.3	562.1
SSE		219145	190708	187748
F (model comparison)			9.99*	5.27*
R^2		0.23	0.33	0.34
adj R^2		0.22	0.31	0.32
AIC		2609.0	2570.6	2567.1
AICC		2609.5	2571.9	2568.6
SBC		2289.8	2270.7	2271.1
CV PRESS		230526	206001	203799

Notes. Entries in the Model columns above Model df are the number of retained predictors in that model where predictors were retained only if their inclusion resulted in a decreased predicted residual error sum of squares (PRESS). For fit measures AIC, AICC, and SBC, lower values indicate better fit. Fit statistics are Akaike information criterion (AIC) (Darlington, 1968), a corrected version of AIC (AICC) (Hurvich & Tsai, 1991), the Sawa Bayesian Information Criterion (BIC) (Sawa, 1978), the Schwarz Bayesian Information Criterion (SBC) (Schwarz, 1978), the predicted residual sums of squares (PRESS), adjusted R^2 (Darlington, 1968). Acronyms are SAS's names for the fit statistics (Cohen, 2006).

Model 5 includes the traditional predictors and school effects. Model 6 adds the PSQ variables (dimension scores). Of the 14 PSQ variables 5 were retained after cross-validation and adjusted R^2 increased from 0.22 to 0.31. All other model fit indices (AIC, AICC, SBC) similarly suggest that the PSQ variables added to the prediction over traditional variables. Model 7 adds the polynomial and interaction terms to capture nonlinearities in the relationship between PSQ and grade point average. By the *F* test and by most of the fit measures (adjusted R^2, AIC, AICC) Model 7 fits the data better than model 6, but by one measure (SBC) Model 6 fits better. The adjusted R^2 difference is rather small from 0.31 to 0.32. A reasonable interpretation of these results is that there is some evidence for nonlinear relationships in the data, and that probably either model could be used to form a best composite that could be used for prediction.

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

Appendix E: Subgroup Differences

Group mean differences were evaluated using the two-tailed independent sample t-test and the effect size measure, Cohen's *d.* Cohen's *d* (Cohen, 1998) is a commonly used measure of effect size for a t-test and has the following interpretation: 0.2 (small effect), 0.5 (moderate effect), and 0.8 (large effect).Table 9 presents the results of subgroup comparisons. The highlighted cells are those with effect sizes larger than 0.20 and a statistically significant t-test. The ones highlighted in green are positive, while the ones highlight in light pink are negative.

TABLE 9. SUBGROUP MEAN DIFFERENCE EFFECT SIZES ACROSS GENDER, RACIAL AND ETHNIC GROUPS

Focal Group **Referent Group**	Female (N=222) - Male (N=218)		Black (N=424) - White (N=249)	
Variable				
Achievement	0.20		-0.01	
Organization	0.33	*	-0.11	
Responsibility	0.17		-0.08	
Abidance	0.28		-0.03	
Self-Discipline	-0.02		-0.12	
Resiliency	-0.46	*	-0.10	
Even-Temperedness	-0.48	*	-0.18	
Leadership	0.04		0.06	
Sociability	0.09		0.07	
Consideration	-0.04		-0.21	
Trust	-0.11		-0.39	*
Artistic Appreciation	0.22		-0.28	*
Curiosity	-0.23		-0.38	*
Creativity	-0.04		-0.15	
Self-regulation	0.20		-0.09	
Engaging with Others	0.07		0.07	
Teamwork	-0.08		-0.32	*
Emotion Regulation	-0.49	*	-0.14	
Innovativeness	-0.02		-0.31	*

Note. Values are Focal - Referent group effect sizes. *Bonferroni adjusted (by outcome) $p < .05$.

Caution should be exercised when interpreting and examining subgroup mean differences. Only a subset of the pilot AMT participants and pilot school students provided background information to allow for subgroup comparison. The representativeness of the samples is limited, so over-generalization of the differences observed should be avoided. Examining the overall pattern of differences are more informative than focusing on particular differences on individual dimensions and composites.

Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.

Authors:
Patrick Kyllonen, Adele Tan

Data Analysts:
Steven Holtzman, Nimmi Devasia

Psychometrician:
Jianbin Fu

Research Project Manager:
Daniel Fishtein

Reviewers:
Daniel F. McCaffrey, Laura S. Hamilton, David Klieger,
Steven C. Pietsch, Kathryn Pedley, Steve Robbins



Copyright © 2025 by ETS. The Eight-Point logo is a trademark of ETS.