Measuring the Power of Learning.®

Research Report
ETS RR–17-01

Investigating Validity Evidence for the *ETS*® Proficiency Profile

Katrina Crotts Roohr

Ou Lydia Liu

Huili Liu

February 2017

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

Investigating Validity Evidence for the *ETS*® Proficiency Profile

Katrina Crotts Roohr,[1] Ou Lydia Liu,[1] & Huili Liu[2]

1 Educational Testing Service
2 Insight Policy Research, Inc.

The *ETS*® Proficiency Profile (EPP), a college-level assessment, has been widely used to evaluate general education student learning outcomes (SLOs) in college. The purpose of this study was to investigate validity evidence for the EPP by evaluating the relationship with outcomes such as student retention, cumulative grade point average (GPA), and degree attainment, and by investigating differential validity across subgroups and cross-sectional learning gains. Three main conclusions were drawn from this study: (a) Students made significant learning gains from freshman to senior year using EPP scores; (b) freshman scores showed modest relationships with cumulative GPA at various points in college and senior scores showed strong relations with final-year cumulative GPA; and (c) differential validity was found across gender, race, and college major when looking at the relationship between EPP scores and first-year and sophomore GPA. Implications of these results are discussed.

Keywords *ETS*® Proficiency Profile; student learning outcomes; higher education; assessment; validity

doi:10.1002/ets2.12127

Student learning outcomes (SLOs) are important skills, attitudes, or competencies that students are expected to acquire at higher education institutions (HEIs; National Institute for Learning Outcomes Assessment [NILOA], 2012). Influences and pressures from statewide governing boards, state mandates, regional and program accreditors, and a drive for accountability have resulted in increased measurement of SLOs across HEIs in the United States (Kuh, Jankowski, Ikenberry, & Kinzie, 2014). SLO assessments such as the *ETS*® Proficiency Profile (EPP) are typically used to satisfy accreditation and accountability requirements, conduct trend analysis, compare students' achievement levels across institutions, and advise students to help them achieve academic success (Educational Testing Service [ETS], 2015).

Examples of important college-level outcomes have traditionally included college grade point average (GPA), enrollment, persistence, and degree attainment (Association of American Colleges and Universities [AAC&U], 2011; Toiv, 2013). These outcomes are used to help measure instructional improvement and student learning (Toiv, 2013). The general knowledge and skills assessed by SLO assessments are also an important part of college-level outcomes. Because both traditional college-level outcomes and SLO assessments can be used for instructional improvement, the relationships between these outcomes can be examined to provide validity evidence of SLO assessments. Validity evidence supports the interpretation of assessment scores by ensuring that the assessment is measuring what it purports to measure (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Thus, this current study intends to investigate the relationship between assessment scores from the EPP, a direct SLO measure of critical thinking, reading, writing, and mathematics, and college-level outcomes such as student retention (i.e., returning to school after first year of college), cumulative GPA, and degree attainment (i.e., obtaining a degree at the institution within 4, 5, or 6 years). Most importantly, this study investigated an important function that SLO assessment could serve: evaluating learning gains between freshmen and seniors.

Corresponding author: Katrina Crotts Roohr, E-mail: kroohr@ets.org

Validity Evidence for Existing Student Learning Outcomes Assessments

Relationships With College-Level Outcomes

According to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), one of the five sources of validity evidence is evidence based on relations with other variables. This includes evaluating test-criterion relationships to see how well particular outcomes of interest measured at the same or later time are related to test scores on an assessment that purports to measure relevant constructs (AERA et al., 2014). Existing research evaluating validity evidence based on test-criterion relationships for SLO assessments has most commonly used college GPA. Research revealed that performance on standardized SLO assessments (i.e., Collegiate Assessment of Academic Proficiency [CAAP], Collegiate Learning Assessment [CLA], and EPP) has been significantly related to college GPA, with small to moderate correlations ranging from .23 to .38 (ACT, 2012; Hendel, 1991; Liu & Roohr, 2013; Marr, 1995; Zahner, Ramsaran, & Steedle, 2012).

Few studies have investigated relationships between SLO assessment scores and other college-level outcomes such as student retention (or persistence) or degree completion. To date, most research evaluating relationships between these outcomes has focused on self-reported SLOs using data from indirect SLO assessments such as the National Survey of Student Engagement (NSSE) and the Community College Survey of Student Engagement (CCSSE). Results for NSSE show some small relationships between level of engagement and higher persistence rates (NSSE, 2010). For the CCSSE, small relationships have also been found between CCSSE score and first to second term and first to second year persistence rates. Additionally, small relationships ranging from .05 to .11 have been found between CCSSE score and degree completion within 3 years (McClenney & Marti, 2006). Despite these existing studies, different relationships may appear when using direct SLO assessments that provide direct evidence of student learning rather than indirect evidence through the use of self-report.

Differential Validity

It is important to establish validity evidence for subgroups to ensure that the interpretation of test scores is comparable (AERA et al., 2014). Differential validity evaluates whether the relationship between an assessment and criterion varies in magnitude by subgroup (Mattern, Patterson, Shaw, Kobrin, & Barbuti, 2008; Young, 2001). Among current SLO assessments, differential validity has been investigated for the EPP across students with English as a second language (ESLs) and non-ESLs. Results showed that the relationship of credit hours completed with EPP scores was the same for both ESLs and non-ESLs (Lakin, Elliott, & Liu, 2012). Although differential validity can provide validity evidence for subgroups, it is important to note that differences in the magnitude of correlation coefficients can also arise from measurement error and may not indicate differences in score meaning, especially if there are mean differences in test scores across the student groups (AERA et al., 2014). As a result, we must use caution when interpreting these results to avoid misleading conclusions.

Using Quartile Comparisons to Evaluate Validity Evidence

Another approach to evaluating validity evidence based on relations with other variables was proposed by Bridgeman, Burton, and Cline (2009). This straightforward method provides an alternative to traditional correlation analyses. The authors argued that correlations might not be useful for conveying information to nontechnical audiences, and they proposed looking at the value of test scores by comparing low- and high-performing students (i.e., dividing students into quartiles based on their performance). Arum, Cho, Kim, and Roksa (2012) used a similar approach for examining post-graduation outcomes using CLA quintiles. Examining differences between low- and high-performing students has the benefit of providing additional information to the validity analyses and is also easier to communicate to higher education stakeholders who may not be measurement experts.

Rationale and Research Questions

Despite the fact that previous studies have provided validity evidence to support the use of existing SLO assessments such as EPP, validation is an ongoing process, and additional evidence should be collected to support the intended uses of test scores based on new data. To date, there have been some studies that have evaluated validity evidence in relation to

other variables such as GPA and credit hours (e.g., Hendel, 1991; Liu & Roohr, 2013; Marr, 1995), and studies that have evaluated differential validity for ELP students (e.g., Lakin et al., 2012). However, additional evidence should be collected to evaluate how other college-level outcomes could be related to EPP performance and whether these relationships vary across subgroups and college majors. Thus, the purpose of this study was to investigate validity evidence for the EPP, a college-level assessment, by evaluating its relationship with college-level outcomes (student retention, cumulative GPA, and degree attainment) and differential validity across gender, race, and college major. We also evaluated a very important function that SLO assessments such as EPP could serve—evaluating learning gains between freshmen and seniors. We addressed the following research questions:

1. What are the cross-sectional learning gains from freshman to senior year of college?
2. What is the relationship of EPP assessment scores with first- to second-year student retention, cumulative GPA, and degree attainment?
3. Is there differential validity among students with varying demographics and college majors?
4. Is the relationship between EPP assessment score and college outcome variables different for low-, middle-, and high-performing students?

These research questions will help contribute to the usefulness, acceptance, and sustainability of EPP for use at HEIs. For instance, these research questions will help to provide additional validity evidence to support the use of EPP scores at HEIs such as the evaluation of trends, internal and external benchmarking, and curriculum and instructional improvement.

Method

Participants

Data for this study were from a large, 4-year public institution in the southern United States. The sample included freshman and senior students who took the EPP from 2007 to 2013. The EPP is commonly administered to both freshmen and seniors within institutions as a way to measure cross-sectional learning gains from freshman to senior year of college. As shown in Table 1, the full sample of freshmen and seniors were at least half female, predominately White, and English-speaking. Non-White students included Asian, Black, Hispanic, and "other." Additionally, freshmen comprised mainly full-time students, and most students were retained from freshman to sophomore year of college. About 50% of freshmen and seniors were in science, technology, engineering, or mathematics (STEM) related majors. Furthermore, students' average composite SAT scores (critical reading plus mathematics) were slightly over 1200 (out of a total score of 1600). ACT scores were converted using the ACT – SAT concordance table (ACT, 2013).

Instrument

EPP, a college-level assessment administered to 550,000 college students at more than 500 institutions nationwide (ETS, 2015), was used to evaluate validity evidence of SLO assessment scores. The EPP measures skills in reading, writing, mathematics, and critical thinking, and it contains questions in three contexts: humanities, social sciences, and natural

Table 1 Freshman and Senior Cohort Demographics

	Freshmen	Seniors
Sample size (*n*)	6954	1109
% Female	51.5	50.4
% White	85.1	84.1
% Speaks better in English	93.5	90.4
% Full time	97.8	–
% STEM major	56.7	52.6
% Retained after 1st year	95.7	–
% Graduated	16.2	87.8
Mean (*SD*) total SAT score[a]	1224.3 (128.6)	1228.1 (139.2)

Notes. STEM = science, technology, engineering, and mathematics. [a]Calculated based on SAT Critical Reading and Mathematics. This also includes ACT scores converted into SAT total using the SAT – ACT Concordance Table (ACT, 2013).

sciences. The standard form of the assessment contains 108 questions and takes 2 hours to complete, and the abbreviated form contains 36 questions and takes 40 minutes to complete (ETS, 2010). Scaled scores for the total range from 400 to 500 and from 100 to 130 for each of the four subscales of reading, writing, mathematics, and critical thinking (ETS, 2010). The reliabilities of the four EPP subscales range from .78 to .84 (ETS, 2010). The abbreviated form was used in this study and is commonly used by institutions because it takes less time to administer as compared to the standard form. The abbreviated form is intended to be used only at the group level, not the individual student level. Reliability for the total score on the abbreviated form is .77 (ETS, 2010).

Data Analyses

To address the first research question, we calculated the mean and standard deviation of EPP total and subscores for both freshmen and seniors. Independent sample *t*-tests and standardized mean differences (i.e., Cohen's *d*) were conducted to evaluate differences in performance across the samples where .20 is small, .50 is moderate, and .80 is large (Cohen, 1988). Additionally, we evaluated the relationship between EPP scores and admissions test scores. It is important to note that the validity of these results is dependent on equal motivational levels in the freshman and senior samples (e.g., Liu, Bridgeman, & Adler, 2012). In an attempt to account for the issue of motivational differences due to the low-stakes nature of the assessment, only those students who completed at least 75% of the assessment were included in the analyses.

Relations With Other Variables

Correlation and regression analyses were conducted to evaluate the relationship of EPP scores with various college-level outcomes. Pearson or point-biserial correlations were analyzed between freshman EPP scores (total and subscores) and first- to second-year student retention, cumulative GPA (first, sophomore, junior, and senior/final year), and degree attainment (within 4, 5, or 6 years). The guidelines developed by Cohen (1988) were used to determine the magnitude of the correlations where .10 is small, .30 is moderate, and .50 is large. Regression analyses were also conducted to see whether freshman EPP total was significantly related to various outcomes when controlling for student demographics, entering academic ability, and college major.

For all regression analyses, independent variables (IVs) included EPP total, admissions test score (i.e., SAT/ACT total), gender, race (White versus non-White), and college major (STEM versus non-STEM). Logistic regression was used to evaluate both student retention (retained or not retained) and degree attainment (graduated or not graduated), whereas linear regressions were used to evaluate cumulative GPA. To evaluate the importance of each of the IVs, dominance analyses were also conducted. Dominance analysis evaluates the importance of each IV "based on comparisons of unique variance contributions of all pairs of variables to regression equations involving all possible subsets of predictors [i.e., IVs]" (Nathans, Oswald, & Nimon, 2012, p. 9). For example, if there were four variables, X_1, X_2, X_3, and X_4, and we were evaluating the dominance of X_1, we could compare X_1 to all possible sets of IVs in a pairwise fashion (e.g., $\{X_2\}$, $\{X_3\}$, $\{X_4\}$, $\{X_2, X_3\}$, $\{X_2, X_4\}$, $\{X_3, X_4\}$, and $\{X_2, X_3, X_4\}$). With five IVs, a total of 32 separate models were analyzed. Using this method, we evaluated general dominance—one IV's overall average across all models is greater than another IV's overall average (Azen & Budescu, 2003). General dominance shows the proportion of importance based on the estimated R^2 and thus the relative importance of each of the IVs in the model.

Similar analyses were conducted for seniors. Correlations were calculated between senior EPP scores and senior/final GPA (i.e., cumulative GPA at the end of a student's senior year, or at the time of graduation) and degree attainment. Additionally, regression analyses were used to evaluate whether senior EPP total was significantly related to college outcomes when controlling for various covariates, and dominance analysis was conducted to evaluate the importance of EPP total.

Differential Validity

Differential validity was evaluated across various subgroups, including gender, race, and college major. Correlations between the various college-level outcomes were conducted separately across subgroups (e.g., separately for males and females). To investigate whether the correlations across subgroups were significantly different from each other, we first used the Fisher z' transformation of r:

$$z' = \frac{1}{2}\left[\ln(1+r) - \ln(1-r)\right]. \quad (1)$$

Once the transformation was completed, statistical differences between z-scores were tested by computing the normal curve deviate (Cohen, Cohen, West, & Aiken, 2003):

$$z = \frac{z'_1 - z'_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}, \tag{2}$$

where n is the sample size for each student group.

Quartile Comparisons

Quartiles for EPP total included high, medium, and low EPP performers (top 25%, middle 50%, and bottom 25%, respectively). Using this breakdown, we examined the percentage of students in the high-, middle-, and low-performing groups in relation to each outcome variable (e.g., the number of students with high EPP scores achieving a first-year GPA of 3.50 or higher). Bar graphs were created to visually represent the quartile comparisons.

Results

Freshman and Senior ETS Proficiency Profile Performance

Freshman students had an average total score of 457.3 ($SD = 17.7$) and average scores of 114.8 ($SD = 6.0$), 121.2 ($SD = 5.7$), 116.6 ($SD = 4.3$), and 117.5 ($SD = 5.8$) on critical thinking, reading, writing, and mathematics, respectively. Senior students performed significantly higher than freshman students on the total (mean = 462.4; $SD = 18.8$), and across subscores. Average senior EPP scores were 116.1 ($SD = 6.3$), 122.0 ($SD = 5.8$), 117.0 ($SD = 4.3$), and 119.4 ($SD = 5.8$) on critical thinking, reading, writing, and mathematics, respectively.

Figure 1 shows the cross-sectional learning gains from freshman to senior year on EPP total and subscores. Standardized mean differences between freshmen and seniors were small on total score ($d = 0.28$) and across subscores ($d = 0.09$ to 0.33). These results indicated that from freshman to senior year of college, students made significant learning gains in critical thinking, reading, writing, and mathematics based on EPP performance, consistent with previous research (e.g., Arum & Roksa, 2011).

To make these learning gains more meaningful, we also looked at differences in freshman and senior college admissions scores. This allowed us to investigate whether the freshman and senior samples have the same academic ability (as measured by college admissions score) upon entering college. Correlations between admissions test score and freshman and senior EPP total were .68 ($p < .001$) and .70 ($p < .001$), respectively. Results showed that the freshman sample had an average college admissions score of 1224 ($SD = 128.6$), and the senior sample had an average admissions score of 1228 ($SD = 139.2$). As shown in Figure 1, no significant differences in admissions score were found between the two samples, suggesting that the two samples had similar ability upon entering college.

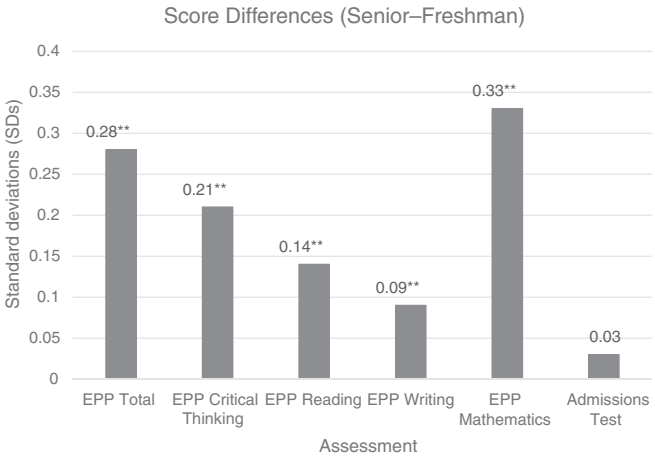

Figure 1 Cross-sectional ETS Proficiency Profile learning gains from freshman to senior year (**$p < .001$).

Table 2 Full Sample and Subgroup Correlations Between ETS Proficiency Profile Score and Grade Point Average (GPA)

		Total		Critical thinking		Reading		Writing		Mathematics	
College-level outcome	*n*	r^a	*z*	r^a	*z*	r^a	*z*	r^a	*z*	r^a	*z*
1st year GPA (Fr)											
All examinees	6954	.28		.21		.21		.13		.19	
Females	3571	.35	4.08**	.26	3.19**	.27	3.89**	.22	2.37*	.26	0.93
Males	3353	.26		.18		.18		.16		.24	
Non-White	948	.31	2.38*	.27	3.00**	.26	2.51*	.19	0.50	.24	2.76**
White	5892	.24		.17		.17		.17		.15	
Non-STEM	3006	.24	−3.40**	.17	−2.76**	.19	−1.86	.17	−1.97*	.16	−2.95**
STEM	3916	.32		.24		.24		.21		.23	
Sophomore GPA (Fr)											
All examinees	2564	.26		.19		.20		.21		.16	
Females	1251	.36	4.02**	.27	3.31**	.28	3.43**	.25	2.35*	.26	1.58
Males	1313	.22		.15		.15		.16		.20	
Non-White	365	.31	1.69	.27	2.05*	.26	1.86	.22	0.46	.22	1.67
White	2177	.22		.16		.16		.19		.13	
Non-STEM	1019	.24	−1.22	.16	−1.49	.19	−0.64	.21	0.00	.14	−1.50
STEM	1543	.28		.22		.21		.21		.19	
Junior GPA (Fr)											
All examinees	1605	.39		.30		.32		.26		.27	
Females	855	.41	−0.14	.29	−0.75	.33	0.11	.24	−1.03	.34	−0.09
Males	750	.42		.32		.32		.29		.34	
Non-White	223	.44	1.58	.34	1.26	.36	1.29	.30	0.99	.37	2.36*
White	1372	.34		.26		.28		.23		.21	
Non-STEM	610	.34	−1.73	.25	−1.72	.29	−0.88	.24	−0.75	.24	−0.83
STEM	995	.42		.33		.33		.27		.28	
Senior/final GPA (Fr)											
All examinees	1473	.26		.22		.21		.16		.14	
Females	883	.28	−0.39	.23	−0.58	.22	−0.08	.15	0.00	.21	−0.29
Males	590	.30		.26		.22		.15		.22	
Non-White	166	.31	1.39	.31	1.74	.24	1.12	.09[b]	−0.55	.24	1.83
White	1259	.20		.17		.15		.14		.09	
Non-STEM	870	.22	−1.60	.19	−1.53	.20	−0.53	.11	−2.29*	.13	−0.42
STEM	603	.30		.27		.22		.23		.15	
Senior/final GPA (Sen)											
All examinees	1109	.41		.32		.34		.29		.26	
Females	559	.42	−0.17	.33	0.12	.34	−0.01	.27	−0.63	.31	−0.13
Males	550	.43		.32		.34		.31		.31	
Non-White	153	.48	2.00*	.37	−0.29	.40	1.68	.30	0.60	.45	3.64**
White	933	.34		.27		.27		.25		.17	
Non-STEM	526	.36	−1.47	.30	−0.51	.33	−0.30	.21	−2.53*	.23	−0.56
STEM	583	.44		.33		.34		.35		.27	

Notes. Fr = freshmen; Sen = seniors; STEM = science, technology, engineering, mathematics.

[a] All correlations significant ($p < .01$). [b] Not significant.

**$p < .01$; *$p < 05$.

Validity Evidence Based on Relations With Other Variables

Relations With Cumulative GPA

Results showed moderate relationships between EPP total and cumulative GPA at different years of college, ranging from .26 to .39 (Table 2). The strongest relationship was between EPP total and junior year GPA. Similar trends were found for EPP subscores, with small correlations ranging from .13 to .22 between freshman EPP subscores and first-year, sophomore, and senior/final GPA. As compared to the relationship with first-year, sophomore, and senior/final GPA, higher correlations were found between EPP subscores and junior GPA, with moderate correlations ranging from .26 to .32.

Regression results in Table 3 showed that freshman EPP total had a significant relationship with cumulative GPA when considering other IVs. Regression equations showed that all IVs together accounted for 18–27% of the variance

in cumulative GPA. In terms of dominance weights, admissions test score had general dominance (i.e., average unique variance across all subset models) over all other IVs when evaluating the relationship with first-year, sophomore, and junior GPA, comprising 35–37% of the estimated R^2. Gender showed the second strongest relationship with first and sophomore GPA (22–34% of R^2) and the strongest relationship with senior/final GPA (35% of R^2). EPP total had the third strongest relationship with first and sophomore GPA (20–21% of R^2), but the second strongest relationship with junior GPA (31% of R^2). Across all regression models, college major had negligible importance (0–2% of R^2).

When evaluating the relationship between senior EPP scores and senior/final GPA, results showed a moderate correlation with total score ($r = .41$) and moderate correlations with subscores, ranging from .26 to .34 (Table 2). Regression analyses showed that senior EPP total had a significant relationship with cumulative GPA when considering other variables (Table 3). All the IVs together accounted for 28% of the variance in senior/final-year GPA. In terms of dominance weights, admissions test score had the strongest relationship (43% of R^2), followed by EPP total (29% of R^2).

Relations With Student Retention

Both point-biserial correlation and logistic regression analyses confirmed that freshman EPP scores did not have a significant relationship with first- to second-year retention for all examinees (Table 4). The finding is likely due to the range restriction, as the institution being analyzed has been very successful in retaining students (e.g., over 90%). College major showed a significant relationship with retention, with STEM majors 26% more likely to drop out of college as compared to non-STEM majors.

Relations With Degree Attainment

Using freshman EPP scores, results revealed small significant point-biserial correlations between 4-, 5-, or 6-year degree attainment and EPP total ($r = .07$), critical thinking ($r = .06$), and writing ($r = .10$); see Table 2. Logistic regression results showed a significant relationship with freshman EPP total (Table 5). Gender also had a significant relationship with degree attainment, with females more likely to graduate than males. Among the other IVs, race and college major had a significant relationship with degree attainment, with White students and non-STEM majors more likely to graduate.

Point-biserial correlation results showed a small significant negative correlation between EPP mathematics and degree attainment (Table 4). This could be a spurious relationship, in that STEM majors may have scored higher in mathematics and may be less likely to graduate due to more challenging courses. To further investigate this, we looked at the breakdown of STEM and non-STEM college majors for students who did not graduate. Overall, the balance between STEM ($n = 74$) and non-STEM ($n = 61$) majors not graduating was fairly equal. Focusing on the STEM majors, the majority of students not graduating were in engineering (28%), genetics and biochemistry (15%), biological sciences (12%), and environmental science (12%). For non-STEM majors, the majority of students not graduating were in business (25%), i.e., accounting and finance, management, economics, and marketing; teacher education in secondary mathematics and chemistry (23%); and political science (13%). These results suggested that the majority of nongraduating seniors were in college majors related to science and mathematics, which could have resulted in the negative relationship between degree attainment and EPP mathematics. Results of the logistic regression analyses found that there were no significant predictors of degree attainment (Table 5).

Differential Validity

Relations With Grade Point Average

Differential validity was evaluated across gender, race, and college major. In terms of the relationship between first-year GPA and freshman EPP score, results showed small to moderate significant correlations for the various subgroups (Table 2). Females showed significantly higher correlations ($r = .22–.35$) as compared to males ($r = .16–.26$) for all EPP scores, except mathematics. Non-White students showed significantly higher correlations ($r = .19–.31$) as compared to White students ($r = .15–.24$) for all EPP scores, except writing. Lastly, STEM majors showed significantly higher correlations ($r = .21–.32$) compared to non-STEM majors ($r = .16–.24$) for all EPP scores, except reading. These results suggest that there are differences in how EPP scores are related to first-year GPA across the various subgroups. Similarly, for sophomore GPA, females showed significantly higher correlations ($r = .25–.36$) than males ($r = .15–.22$) for all EPP scores,

Table 3 Grade Point Average (GPA) Linear Regression and Dominance Weights Analyses

	Freshman sample												Senior sample		
	1st year GPA			Sophomore GPA			Junior GPA			Senior/final GPA			Senior/final GPA		
	β(SE)	*DW*	*%R²*	*β(SE)*	*DW*	*%R²*	*β(SE)*	*DW*	*%R²*	*β(SE)*	*DW*	*%R²*	*β(SE)*	*DW*	*%R²*
Intercept	−.321 (.200)			.257 (.296)			−.493 (.301)			1.043** (.303)			−.094 (.346)		
EPP total	.004** (.001)	.041	21%	.003** (.001)	.036	20%	.006** (.001)	.083	31%	.002* (.001)	.032	13%	.004** (.001)	.082	29%
Admissions Test score	.001** (.000)	.072	37%	.001** (.000)	.065	35%	.001** (.000)	.098	37%	.001** (.000)	.074	29%	.001** (.000)	.121	43%
Gender (male)	−.312** (.014)	.052	27%	−.305** (.021)	.063	34%	−.257** (.022)	.058	22%	−.311** (.023)	.088	35%	−.227* (.026)	.046	16%
Race (White)	.217** (.022)	.025	13%	.143** (.031)	.016	9%	.149** (.032)	.026	10%	.310** (.037)	.057	23%	.165* (.038)	.030	11%
College major (STEM)	−.085** (.015)	.004	2%	−.075** (.022)	.004	2%	−.040 (.023)	.002	1%	−.018 (.024)	.001	0%	.001 (.026)	.001	0%
R	.440			.429			.517			.502			.530		
R^2	.193			.184			.267			.252			.281		

Notes. DW = dominance weights. All DWs add up to the R^2; $\%R^2$ = calculated by dividing the DW by the R^2. Indicates the proportion of contribution towards the estimated R^2 for that independent variable.

**$p < .01$; *$p < 05$.

Table 4 Full Sample and Subgroup Correlations Between ETS Proficiency Profile Score and Retention and Degree Attainment

		Total		Critical thinking		Reading		Writing		Mathematics	
College-level outcome	*n*	*r*	*z*	*r*	*z*	*r*	*z*	*r*	*z*	*r*	*z*
Retention (Fr)											
All examinees	6954	.01		.01		.00		.03		.01	
Females	3579	.02	−0.07	.02	−0.83	−.01	0.54	.03	−0.12	.00	0.56
Males	3375	.01		.00		.01		.03		.01	
Non-White	954	−.02	−1.07	−.01	−0.68	−.03	−0.77	.01	−0.70	−.03	−1.03
White	5916	.02		.01		.002		.03*		.01	
Non-STEM	3013	−.03	−3.31**	−.05*	−3.96**	−.03	−2.09*	−.01	−2.47*	.01	−0.36
STEM	3939	.05**		.05**		.02		.05**		.01	
Degree attainment (Fr)											
All examinees	1362	.07*		.06*		.05		.10**		−.02	
Females	799	.07*	0.28	.06	0.44	.06	−0.41	.06	0.85	.02	0.17
Males	563	.09*		.08		.04		.11**		.03	
Non-White	170	−.02	0.83	−.03	0.97	−.06	1.23	.09	−0.14	−.02	−0.33
White	1144	.05		.06		.04		.07		−.04	
Non-STEM	865	.04	−1.63	.05	−0.98	.03	−0.70	.05	−2.01*	−.02	−0.59
STEM	497	.13**		.10*		.08		.16**		.02	
Degree attainment (Sen)											
All examinees	1083	−.03		−.02		.01		−.03		−.06*	
Females	559	−.09*	0.65	−.06	0.67	−.01	−0.45	−.06	0.48	−.13**	1.26
Males	550	−.05		−.02		−.04		−.03		−.06	
Non-White	153	.01	−0.97	.11	−1.84	−.03	0.12	.003	−0.59	−.08	−0.18
White	933	−.07*		−.05		−.02		−.05		−.09**	
Non-STEM	526	−.09*	0.89	−.09*	1.81	−.01	−0.47	−.07	0.71	−.10*	0.14
STEM	583	−.04		.02		−.04		−.03		−.09*	

Note. Fr = freshmen; Sen = seniors; STEM = science, technology, engineering, mathematics.
**$p < .01$; *$p < 05$.

Table 5 Degree Attainment Logistic Regression

	Freshmen		Seniors	
	4-, 5-, or 6-year degree attainment		Degree attainment	
	β(SE)	*Exp(β)*	*β(SE)*	*Exp(β)*
Intercept	4.54 (0.97)		3.73 (2.88)	
EPP total	−0.01* (0.003)	0.99	−0.001 (0.01)	1.00
Admissions test score	−0.001 (0.00)	1.00	−0.05 (0.21)	0.95
Gender (male)	−0.43* (0.07)	0.65	0.05 (0.32)	1.05
Race (White)	0.57* (0.12)	1.76	−0.001 (0.001)	1.00
College major (STEM)	−1.06* (0.07)	0.35	0.01 (0.21)	1.01

*$p < .001$.

except mathematics, and non-Whites ($r = .27$) showed a significantly higher correlation as compared to Whites ($r = .16$) on critical thinking. There were no significant differences between STEM and non-STEM majors.

For junior GPA, correlations were small to moderate. In terms of differences in magnitude, the only significant difference was found between non-Whites ($r = .37$) and Whites ($r = .21$) for EPP mathematics. For senior/final GPA, all correlations were significant, except for the correlation for non-Whites with writing. Differences in the strength of the correlation were found on writing, with STEM majors showing significantly higher correlations ($r = .23$) as compared to non-STEM majors ($r = .11$). These results suggested that there is minimal differential validity when evaluating the relationship between freshman EPP scores and junior or senior/final GPA.

For the relationship between senior/final GPA and senior EPP scores, all correlations across subgroups were small to moderate and significant. No differences were found between males and females; however, stronger correlations were found for non-Whites on EPP total ($r = .48$) and mathematics ($r = .45$) as compared to White students ($r = .34$ and .17,

respectively). Across college majors, a stronger relationship was found for STEM majors on writing ($r = .35$) as compared to non-STEM majors ($r = .21$).

Relations With Student Retention

When evaluating the relationship between EPP scores and student retention across subgroups, results showed no significant correlations across gender and race (see Table 4). For college major, significant correlations were found between student retention and EPP total, critical thinking, and writing; however, correlations were extremely low ($r < .10$). Additionally, significant differences in the strength of these relationships were found for all scores except mathematics, with stronger relationships found for STEM majors as compared to non-STEM majors (differences ranged from .05 to .10). These results suggest some differential validity across college majors.

Relations to Degree Attainment

Results evaluating the relationship between freshman EPP score and degree attainment showed very small, but significant, relationships between EPP total and degree attainment for females ($r = .07$), males ($r = .09$), and STEM majors ($r = .13$); see Table 4. The difference in correlation magnitude was not significantly different across males and females. Small, significant correlations were also found between critical thinking and degree attainment for STEM majors ($r = .10$) and with writing for males ($r = .11$) and STEM majors ($r = .16$). Results also showed differential validity across college major when evaluating the relationship between degree attainment and freshman EPP writing (STEM $r = .16$; non-STEM $r = .05$).

Using senior EPP score, results showed small negative correlations between EPP total and degree attainment for females ($r = -.09$), Whites ($r = -.07$), and non-STEM majors ($r = -.09$); see Table 4. Small negative correlations were also found with critical thinking for non-STEM majors ($r = -.09$) and with mathematics for females ($r = -.13$), White students ($r = -.09$), non-STEM ($r = -.10$), and STEM majors ($r = -.09$). There were no significant differences in the relationships across STEM majors; however, as previously discussed, this was likely due to the fact that the non-STEM majors consisted of students enrolled in mathematics-related college majors. Therefore, these students likely showed a similar pattern to STEM students.

Quartile Comparisons

Focusing on the significant results from the regression analyses, we further analyzed the relationship between EPP total and college-level outcomes using quartile comparisons. Figure 2a – d shows the relationship between freshman EPP total and cumulative GPA across various years of college. Results showed that high EPP performers were 2 – 3 times more likely to achieve a 3.50 GPA or higher as compared to low performers. Similar trends were found for senior EPP total when evaluating the relationship with senior/final GPA (Figure 3), with high EPP performers 3.1 times more likely to achieve a 3.50 or higher GPA compared to low performers.

Conclusions and Discussion

The purpose of this study was to investigate validity evidence and differential validity of the EPP, an SLO assessment, in terms of the relationship with college-level outcomes including cumulative GPA, student retention, and degree attainment. Additionally, we also evaluated cross-sectional learning gains. Results showed significant learning gains from freshman to senior year using EPP scores. Overall, EPP scores showed consistent relationships with cumulative GPA; however, results showed some differential validity across gender, race, and college major when evaluating the relationship with first-year and sophomore GPA.

Validity Evidence for ETS Proficiency Profile Scores

Results indicated that freshman EPP total scores consistently have modest validity evidence when evaluating the relationship with cumulative GPA, and senior EPP total scores have fairly strong validity evidence when evaluating the relationship with senior/final GPA. Additionally, EPP total consistently had a significant relationship with these college outcomes when

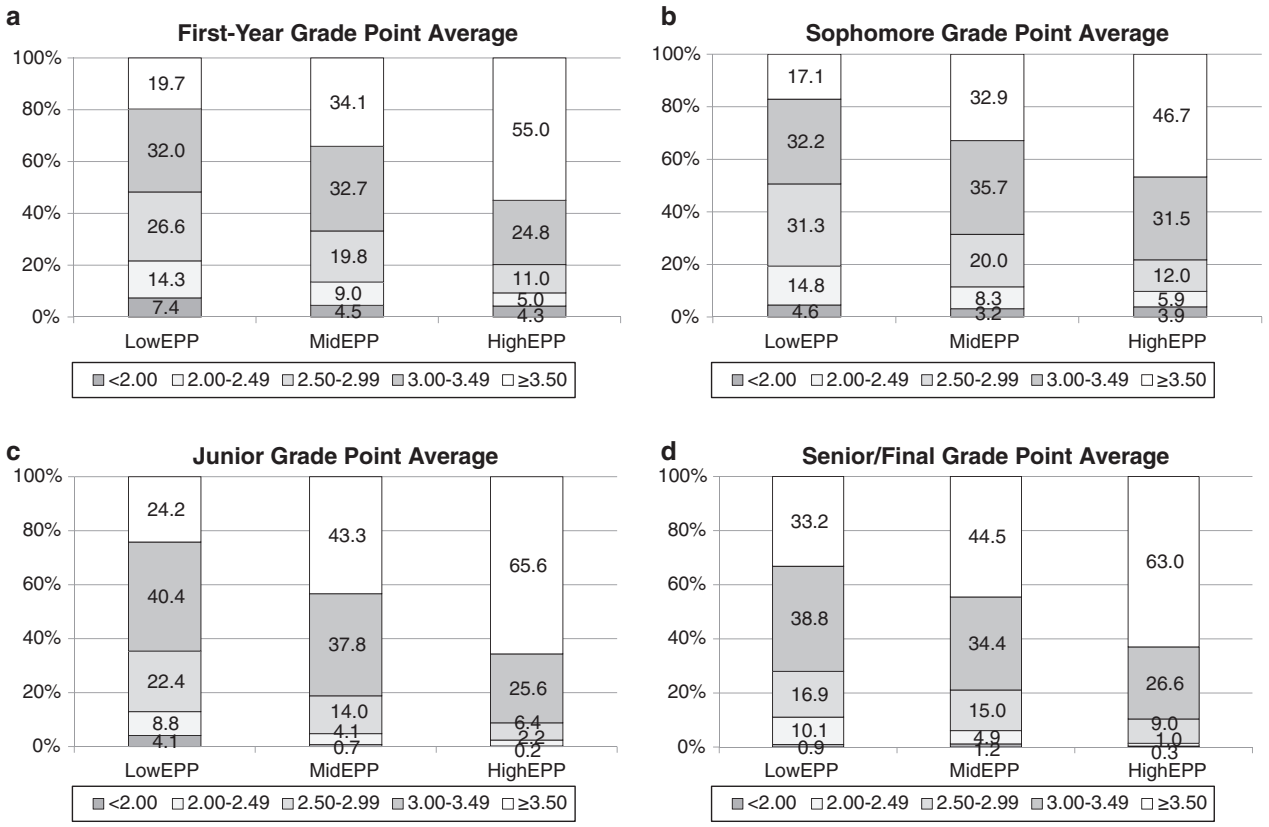

Figure 2 (a)–(d) Freshman sample quartile comparisons. These graphs show the relationship between high, medium, and low ETS Proficiency Profile performers (top 25%, middle 50%, and bottom 25%, respectively) and cumulative grade point average.

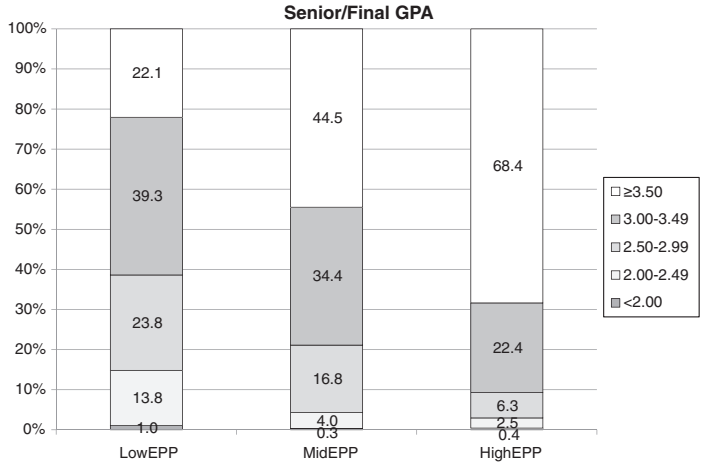

Figure 3 Senior sample quartile comparison. These graphs show the relationship between high, medium, and low ETS Proficiency Profile performers (top 25%, middle 50%, and bottom 25%, respectively) and senior/final cumulative grade point average.

considering other covariates. These results are consistent with previous studies that have investigated the relationship between SLO assessment scores (on the CAAP, EPP, and CLA+) and GPA (e.g., ACT, 2012; Hendel, 1991; Zahner et al., 2012). In terms of dominance weights, EPP total comprised 21–31% of the estimated R^2 when evaluating the relationship with first-year, sophomore, or junior GPA; however, when evaluating the relationship with senior/final GPA, it comprised only 13% of the estimated R^2. Although EPP was not the strongest predictor of cumulative GPA, these results show that EPP total does play an important role in evaluating the relationship with cumulative GPA.

In terms of student retention, no significant relationships were found. A possible reason for this result could be due to range restriction in the criterion variable with this sample, as the institution included in the analysis has a very high first to second year retention rate (e.g., over 90%). The results may be different when another sample with a lower retention rate is analyzed.

For degree attainment, significant correlations were found using freshman EPP critical thinking and writing when evaluating the relationship with degree attainment, suggesting that students with higher critical thinking and writing skills in their freshman year of college are more likely to graduate within 6 years. Critical thinking and writing skills have consistently been identified as important SLOs by institutions (e.g., AAC&U, 2011). These results support the importance of these skills for graduation. Interestingly, these same trends were not found when evaluating the relationship between senior EPP total and degree attainment. In fact, results showed a negative relationship between senior EPP mathematics performance and degree attainment. This may be a spurious relationship, in that college majors related to science and mathematics may have scored higher in mathematics and may be less likely to graduate due to more challenging courses. To further investigate this, we examined non-STEM and STEM major students who did not graduate. Results showed that the majority of non-STEM majors were mathematically related (e.g., teacher education in secondary mathematics, business and finance majors).

Differential Validity by Gender, Race, and College Major

Differential validity analyses revealed that freshman EPP results had stronger relationships with first-year GPA for females, non-Whites, and STEM majors. Similar results were found for sophomore GPA across gender. For junior or senior/final GPA, non-Whites had a stronger relationship between EPP mathematics and junior GPA, and STEM majors had a stronger relationship with EPP writing and senior/final GPA. Gender results were consistent with studies evaluating differential validity between admissions test scores and first-year GPA. Previous research has shown that admissions test scores are more related to first-year GPA for females than males (Mattern et al., 2008; Young, 2001). For minority groups, previous research is mixed. Some research has found that admissions test scores are more related to first-year GPA for White students (e.g., Mattern et al., 2008). However, Young (2001) indicated that some studies (especially those with small sample sizes) showed opposite results, with minority students having higher correlations between admissions test score and first-year GPA as compared to White students. Because our sample of non-White students was small, this could explain why we were seeing higher non-White correlations than White student correlations when looking at the relationship between EPP score and GPA.

Results for STEM majors were also in line with previous research. Steedle and Bradley (2012) found that students in STEM majors such as natural science and technology engineering, and math majors, performed highest when taking the CLA. Given that these students are typically high performers on SLO assessments, SLO assessments may be more likely to be related to certain college-level outcomes such as GPA for those particular college majors. Similarly, this previous research supports the stronger relationship with student retention for STEM majors over non-STEM majors.

Benefit of Quartile Comparisons

Bridgeman et al. (2009) argued that multiple correlations, although convenient, may not be the easiest for nontechnical audiences to interpret and proposed an alternative approach using quartile comparisons. Using this method, we presented relationships between various levels of EPP performance and college-level outcomes using interpretable graphical displays. These displays clearly show the relationship between freshman EPP score and cumulative GPA. This visual information may be useful for institutions to identify students who might be struggling academically. Using quartiles, an institution may be able to better evaluate the distribution of EPP scores across students and visually perceive the relationships. Future research could also consider developing various graphs across college majors or demographic groups to provide additional information for instructional improvement.

Implications, Limitations, and Future Research

This study has important implications for institutions using the EPP. Findings from this study showed the strong relationship between freshman EPP score and first-year and sophomore GPA. Results also showed that the relationship is

stronger for females, non-Whites, and STEM majors. Results of this study may also generalize to other general-skills SLO assessments (such as the CLA+ and CAAP), as previous research showed that EPP scores correlated strongly (e.g., $r > .70$) with other standardized tests measuring similar constructs (e.g., Klein et al., 2009).

Other important implications can be drawn from the results of the differential validity analyses. Results for the relationship between EPP score and first-year GPA and retention for STEM and non-STEM majors provide important information to stakeholders. These results can inform stakeholders that STEM majors who perform high on EPP may be more likely to obtain a higher first-year GPA and be retained from freshman to sophomore year. Given that STEM majors may have a more difficult course load, freshman EPP scores could be useful for institutions in identifying students who may drop out. Similarly, differential validity results across gender and race showed that EPP scores may not be as related to GPA for males or White students, which should be considered when evaluating EPP freshman scores. Future research should also consider other demographic groups, such as international students.

The study has limitations that may have impacted the generalizability of the findings. First, we were unable to control for student motivational levels using item response time data or a motivational survey. Instead, we were only able to identify those students who did not complete at least 75% of the assessment. Because there were potentially unmotivated students in the sample, some of the correlations may be lower than the true relationships between EPP scores and college-level variables. Additionally, because freshman and senior students may have had different motivational levels, the amount of learning gains from freshman to senior year may also not be representative of the true growth in that institution. In relation to student learning gains, future research should also consider using a longitudinal method rather than a cross-sectional method.

Another limitation is that the data were from only one institution, which may not be representative of HEIs in the United States. As discussed earlier, this particular institution has a very high retention rate, which is certainly not representative of all U.S. HEIs. As a result, EPP scores did not show a significant relationship with retention due to range restriction in the criterion variable. In terms of future research, the analyses should be replicated with samples from additional institutions to see if the results remain the same. Additionally, future research should also consider evaluating retention from second to third year of college and from third to fourth year of college. Lastly, additional criteria should be gathered to expand the relationship of SLO scores of key postcollege outcomes.

Acknowledgments

The authors would like to thank Chen Li for her assistance in conducting the dominance analyses. The authors would also like to thank Kri Burkander and Jennifer Bochenek for their assistance in formatting and editing the paper and references.

References

ACT, Inc. (2012). *ACT CAAP technical handbook 2011–2012*. Iowa City, IA: CAAP Program Management.

ACT, Inc. (2013). *ACT–SAT concordance: A tool for comparing scores*. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/reference.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arum, R., Cho, E., Kim, J., & Roksa, J. (2012). *Documenting uncertain times: Post-graduate transitions of the academically adrift cohort*. New York, NY: Social Science Research Council.

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago Press.

Association of American Colleges and Universities. (2011). *The LEAP vision for learning: Outcomes, practices, impact, and employers' view*. Washington, DC: Author.

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods 8*(2), 129–148. doi:10.1037/1082-989X.8.2.129

Bridgeman, B., Burton, N., & Cline, F. (2009). A note on presenting what predictive validity numbers mean. *Applied Measurement in Education, 22*(2), 109–119. doi:10.1080/08957340902754577

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J., Cohen, P. S., West, G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Educational Testing Service. (2010). *ETS Proficiency Profile user's guide*. Princeton, NJ: Author.

Educational Testing Service. (2015). *ETS Proficiency Profile*. Retrieved from http://www.ets.org/proficiencyprofile/about

Hendel, D. D. (1991). Evidence of convergent and discriminant validity in three measures of college outcomes. *Educational and Psychological Measurement, 51*, 351–388. doi:10.1177/0013164491512008

Klein, S., Liu, O. L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., … Steedle, J. (2009). *Test Validity Study (TVS) report*. New York, NY: Collegiate Learning Assessment.

Kuh, G. D., Jankowski, N., Ikenberry, S. O., & Kinzie, J. (2014). *Knowing what students know and can do: The current state of student learning outcomes assessment in U.S. colleges and universities*. Urbana: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Lakin, J. M., Elliott, D. C., &. Liu, O. L. (2012). Investigating ESL students' performance on outcomes assessments in higher education. *Educational and Psychological Measurement, 72*(5), 734–753. doi:10.1177/0013164412442376

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41*(9), 352–362. doi:10.3102/0013189X12459679

Liu, O. L., & Roohr, K. C. (2013). *Investigating ten-year trends of learning outcomes at community colleges* (Research Report No. RR-13-34). Princeton, NJ: Educational Testing Service. 10.1002/j.2333-8504.2013.tb02341.x

Marr, D. (1995). *Validity of the academic profile*. Princeton, NJ: Educational Testing Service.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential predictive validity and prediction of the SAT* (Research Report No. 2008-4). New York, NY: The College Board.

McClenney, K. M., & Marti, C. N. (2006). *Exploring relationships between student engagement and student outcomes in community colleges: Report on validation research* (Working Paper). Austin, TX: The Community College Survey of Student Engagement.

Nathans, L. L., Oswald, F. L., & Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research & Evaluation, 17*(9), 1–19.

National Institute for Learning Outcomes Assessment. (2012). *Providing evidence of student learning: A transparency framework*. Retrieved from http://www.learningoutcomeassessment.org/TFComponentSLOS.htm

National Survey of Student Engagement. (2010). *Validity: Predicting retention and degree progress*. Retrieved from http://nsse.iub.edu/pdf/psychometric_portfolio/Validity_RetentionAndDegreeProgress.pdf

Steedle, J. T., & Bradley, M. (2012, April). *Majors matter: Differential performance on a test of general college outcomes*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, British Columbia.

Toiv, B. (2013). *Top research university expanding efforts to assess, improve undergraduate student learning.* Retrieved from http://www.aau.edu/WorkArea/DownloadAsset.aspx?id=14849

Young, J. W. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis* (College Board Research Report No. 2001-6). New York, NY: The College Board.

Zahner, D., Ramsaran, L. M., & Steedle, J. T. (2012). *Comparing alternatives in the prediction of college success*. New York, NY: Council for Aid to Education.

Suggested citation:

Roohr, K. C., Liu, O. L., & Liu, H. (2016). *Investigating validity evidence for the ETS® Proficiency Profile* (Research Report No. RR-17-01). Princeton, NJ: Educational Testing Service. https://dx.doi.org/10.1002/ets2.12127

Action Editor: Donald Powers

Reviewers: Yigal Attali and Brent Bridgeman