Who Needs ETS? How to Create a Psychometric Assessment Test

Steven A. Frankforter
Winthrop University

Terri L. Guidry
Winthrop University

In this paper, we suggest that colleges of business might create their own reliable assessment tests as an alternative to the purchase of the Educational Testing Service (ETS) Major Field Tests. First, we review the advantages and disadvantages of employing ETS or other standardized tests. Next, we describe the process of creating a test based on psychometric principles. Finally, we cover how to process the resulting data with exploratory factor analysis to determine possible useable questions that are then evaluated for reliability with Cronbach's Alpha statistics.

WHO NEEDS ETS? HOW TO CREATE A PSYCHOMETRIC ASSESSMENT TEST INTRODUCTION

Educational Testing Service (ETS) is the creator and distributor of a variety of assessment tests that universities use for program assessment. ETS test instruments are psychometric in design, objectively measuring skills and knowledge. They have statistically-confirmed reliability measures that collegiate degree programs often used for assessment. ETS offers Major Field Tests for many traditional Arts & Science areas such as biology and political science. They also sell tests for programs frequently offered in colleges of business: tests for undergraduate business, economics, computer science, and one graduate-level test, for MBA. However, colleges might attempt to create their own reliable psychometric assessment tests as an alternative to the purchase of ETS Major Field Tests.

In this paper, we will review the advantages and disadvantages of ETS Major Field Tests and for creating an institution-controlled assessment test. We also describe how to create a reliable institution-controlled psychometric assessment test using SPSS (Statistical Package for Social Sciences). While the focus of this paper will be on business program assessment, the topics are certainly relevant in assessing any academic program.

ADVANTAGES AND DISADVANTAGES OF USING ETS TESTS FOR ASSESSMENT

Advantages of Using ETS for Assessment

The use of ETS Major Field Tests offers numerous advantages. ETS Major Field Tests and other standardized tests invariably have high reliability. For example, the reliability coefficient that ETS reports for the Major Field Test in business is .89 (ETS, 2011a). Next, standardized tests are frequently psychometric-focused, with high perceived acceptance and legitimacy (Banta, ed., 2011), high reliability

(Cangelosi, 1990; Walvoord, 2010). They also benefit from demonstrated construct validity in their assessment measures (Messick, 1993). Standardized tests are readily available and offer benchmark data against which the performance of local students may be compared. (Banta, 2007). Last, they provide summative assessment to document student learning at program exit or associated with a capstone experience, providing evidence of student ability at graduation. (Dwyer, ed., 2008).

Disadvantages of Using ETS for Assessment

ETS Major Field Tests possess narrowly-defined content areas, employing course-grained construct subscore measures which may not align well with an institution's student learning outcomes (Dwyer, ed., 2008), especially with multidisciplinary programs (Walvoord, 2010), making face validity suspect (McMillan, 2008). Additionally, ETS and other standardized tests may not yield actionable data, especially when intended for accountability purposes (Banta, ed., 2011). There are faculty issues as well. Being uninvolved in their creation, faculty are less likely to use the results of standardized tests (Banta ed., 2007). Further, standardized tests do not necessarily contain all essential content areas required by either state mandates or accreditation standards (Banta ed., 2007).

Consequently, standardized tests may not be effectual tools for pinpointing problems and generating solutions. At best, they should be one of several assessment sources (Walvoord, 2010), enriched through multiple instruments providing fine-grained information (Banta, 2002). For example, if a program's mission and student learning goals emphasize content areas not covered in the ETS Business Test, such as leadership, then exclusively reliance on the results of that test is dubious.

Administering an ETS Major Field Test requires a significant commitment of class time, requiring two or more hours to complete and as a summative assessment, student motivation to perform can be suspect (Banta, 2007), which may negatively affect test scores. Finally, when stakes are high, such as with an upcoming accreditation review, faculty may be tempted to teach to a standardized test (Banta, 2007).

Another limitation is that ETS-reported benchmarks are problematic. They are not true national averages. Rather, they are averages of those select schools who have chosen to administered the test frequently enough to be included in ETS norms (Dwyer, ed., 2008). Standardized tests focus on summative assessment, rather than on the institution's contributions to developing student abilities (Ewell, 1984). They also make inter-institutional comparisons difficult because student baseline skills are not measured, but are extremely important in determining an educational institution's value-added (Pike, 2006). Further, ETS' testing administration guidelines make it difficult to track the progress of individual students over time in pretest-posttest situations because it's primarily intended as summative assessment of students approaching graduation http://www.ets.org/mft/about/ (ETS, 2011b).

Lastly, ETS testing can be expensive. The cost of the ETS Major Field Test for Business is at least $24 per student (http://www.ets.org/mft/pricing/), while an institution-created assessment test may pose a reduced out-of-pocket expense and would be nonrecurring (ETS, 2011c). For example, a program administering the ETS Business Test to 300 students per year would incur an annual cost of at least $7,200. However, the out-of-pocket cost of employing an institution-controlled assessment test could be considerably less. However, the piloting and refining of an institution-created assessment test may consume material amounts of time and effort. When taking into account both explicit and implicit costs and benefits, business schools might consider creating institution-controlled assessment instruments with statistically-verifiable reliability measures.

Last, the use of standardized tests may reflect a desire to minimize effort in managing assessment, which may end after data are collected, and not be subject to interpretation or evaluation. Another problem is that such lack of investment in assessment may be reflected in diminished faculty training in evaluating and interpreting the data (Lyman, 1986). Through no fault of ETS, the availability of a standardized test may deter the development of other assessment mechanisms, such as surveys and focus groups (Banta, Jones, & Black, 2009). Last, standardized tests consume scarce resources and may not pinpoint problem areas (Walvoord, 2010).

ADVANTAGES AND DISADVANTAGES OF CREATING INSTITUTION-CONTROLLED ASSESSMENT TESTS

Advantages of Creating Institution-Controlled Assessment Tests

Standardized tests are primarily summative in design (Dwyer, ed., 2008). Whereas, if an institution created its own assessment tests, they could be formative, enhancing teaching and learning within a program. Formative assessment improves student learning ex-post, while summative assessment documents student learning and it's often used with little or no emphasis on using results to improve learning. Course-embedded summative assessment can become a critical component of an effective assessment system (McMillan, 2008).

Incoming student ability is the largest predictor of any outcome (Pascarella & Terenzini, 1991; Willingham, Young, & Morris, 1987). However, the primary use of ETS Major Field Exams is to serve as an exit assessment. If an institution can create and use its own assessment instruments to establish baseline student performance, and the tests are designed to measure skill areas that match school-defined learning targets in a pretest posttest design, it is possible to determine student growth in important skill areas (Hoover, Giambatista, Sorenson, & Boomer, 2010), establishing an academic program's value added through repeated administrations of common tests (Banta ed., 2007). Additionally, employing institution-created tests in a longitudinal design allows for comparisons across subject matter scores that can be useful (Tuckman, 1985). Numerous institution-created assessment instruments allows for the creation of a comprehensive database that reports on the learning progress of individual students. Data can be aggregated across students and courses to evaluate course and program effectiveness (Banta ed., 2007).

Rather than designing learning goals to fit those embedded in standardized tests, the development of institution-controlled assessments helps in the creation of unique learning goals (Cangelosi, 1990). This enables assessment instruments to be designed to closely match an institution's needs.

An important advantage of standardized tests is that they are invariably constructed with a strong scientific foundation with assurance of validity and reliability. However, these advantages can also be achieved in institution-created assessment instruments (Dwyer, ed., 2008). An institution can develop assessment tests that have predicative validity, content validity, construct validity, and are reliable (Nunnally, 1967). Creating in institution-controlled assessment test with established reliability measures will yield consistent results regardless of when the assessment occurs or who does the scoring (Perkins, 1999). With skill and effort, an institution can create assessment instruments that achieve quality comparable to those of standardized tests (Banta, 2011).

Faculty can be impediments to the implementation of assessment. By engaging faculty through assessment in their disciplines, faculty engagement and buy-in can be achieved (Banta ed., 2007). Building local support and creating a culture of assessment is enhanced when local measures are constructed with faculty involvement (Banta & Blaich, 2011).

Managing the out-of-pocket assessment costs may be in important factor in moving away from standardized tests. One can design high-quality valid and reliable tests that can be efficiently reused (Cangelosi, 1990).

Disadvantages of Creating Institution-Controlled Assessment Tests

The creation of a test alone will not ensure a sound assessment system. The system must engage faculty and governance, faculty development systems, and campus leadership to be effectual (Banta & Blaich, 2011). Additionally, it requires considerable time and effort to develop sound assessment instruments and the institution must be capable of this investment before it begins.

After an academic institution evaluates all the advantages and disadvantages of either using ETS Major Field Tests or creating their own instruments for assessment, it should select the approach that best fits its situation. If the decision is to proceed with creating its own assessment test, we describe the process of how to do so in the next section.

DESIGNING ASSESSMENT TEST CONTENT

Creating an assessment test begins with deciding on the necessary coverage areas, which may be driven by a mission, stakeholder needs, and/or accreditation requirements. For example, AACSB requires that coverage be determined from a school's mission statement, but generally offers suggestions, stating that "traditional business subjects" should be covered (AACSB International, 2010: p. 8). The ETS Major Field Test for undergraduate business programs includes 120 questions covering the following subjects: accounting (18 questions), economics (16 questions), management (18 questions), quantitative business analysis (13 questions), information systems (12 questions), finance (16 questions), marketing (16 questions), legal & social environment (12 questions), and international issues (12 questions, shared with other content areas) (ETS, 2011d). Assessment requirements outside these specific areas cannot easily be met with ETS tests. ETS does permit the addition of 50 additional items and an institution could create and administer additional questions among those 50. However, there may be doubtful advantage in doing this. Lengthening the testing period to accommodate additional questions may elongate the testing period beyond the reasonable or available classroom time. Hence, when a college of business attempts to create an assessment test, the process begins with deciding which particular academic areas the mission prioritizes.

The next step is to collect or create the test questions. The questions must be administered as true/false or multiple choice so they can be easily graded and converted to a value of 0 or 1 to indicate "correct" or "incorrect" (Raykov & Marcoulides, 2010). Test bank questions might be used. However, test bank questions might be of uneven quality, inconsistent, or too focused on a topic found only in a particular book. For these reasons, we suggest using questions that reflect standard accepted knowledge that will likely endure over time. Then, as faculty and textbooks change, the questions on the assessment test will continue to be relevant. Creating original questions is also a possibility. If using original questions, the issue of enduring relevance must still be resolved because curriculum evolves over time.

Last, the number of questions per subject must be decided. One should begin with a significant number of questions because the process of establishing construct reliability will lead to the discarding of some questions. However, the fewer the questions used, the greater the possibility that useable results do not materialize because of low Cronbach's Alpha statistics (Nunnally, 1967).

Piloting Assessment Tests

We propose that each knowledge area included in assessment testing have its own, stand-alone test or subscale developed. Business schools may note AACSB requires that "the assessments provide the school with the assurance measures needed to ascertain whether the school's learning goals are being met" (AACSB International, 2010: p. 66). Combining scores of all the knowledge areas into an overall performance measure, and then focusing only on that total, will obscure feedback information relevant to individual knowledge areas. For example, in a situation where overall test results show an increasing trend because accounting scores rose, obscures the problem that other scores, like economics, may have declined. Therefore, we suggest that each knowledge area assessed with a stand-alone test, be evaluated apart from other subject areas tested. Additional benefit is realized when there are changes in programs that require the addition of knowledge areas, because new tests may be developed and then added to the array of those previously created.

In keeping with accepted psychometric theory, the next step in the process of test construction is trying out a form of the proposed instrument or piloting the test with a sample of the intended population. (Raykov & Marcoulides, 2010). Piloting assessment tests can be a time-consuming process. For example, this could include the time that students expend in answering questions, faculty time to oversee questions administration, and administrative time in collating and processing data yielded from testing. To insure best results, where students are properly motivated and their performance monitored, we suggest administering test questions under classroom conditions with pencil & paper, with answers indicated on computer-scanned forms to yield data that can be input and managed electronically. Alternately, online exams could be used in a classroom or lab setting. Of great concern is this question: how much time can

reasonably be committed to the piloting process? The less the piloting process utilizes class time, the more faculty and students will cooperate. The greater the total number of questions piloted with students, the greater the time that must be committed to piloting. Simplifying the tremendous amount of data contained in a battery of tests can be achieved using exploratory factor analysis. The data from test items can be understood in terms of individual variations along a small number of dimensions, called factors. (Hunt, 2011) This issue will be discussed more completely when we cover the use of exploratory factor analysis.

One last concern is the decision of whether or not to pilot the test with students who have not been introduced to a knowledge area (e.g., incoming freshmen). Doing so infuses a high level of randomness in answers. As a result, the standard error would probably rise, increasing the possibility of unusable results.

Statistically Evaluating Assessment Tests

After piloting, the process of refining reliable assessment instruments begins. The first step will be to prepare the data file of graded student answers arising from the piloted tests. The data collected will be stored as a text file with the items and record identifiers as columns, the rows correspond to individuals completing the test and the columns correspond to the questions administered. Import the file into an editable form using Microsoft Excel or ACCESS. Convert text responses (T/F or A,B,C,D,E) to numerical scores that identify answers as either correct or incorrect, coded as 1 or 0. For example, converting the data file to reflect a 1 for each correct response, then summing the number of 1s for an individual student would yield his/her total number of correct answers. After recoding data into a numerical format, statistical analysis may begin with one of many available statistical software applications, such as SPSS or SAS (Statistical Analysis System). Examples of analysis employing SPSS follows.

Exploratory Factor Analysis

Exploratory factor analysis is a key statistical tool that may be applied in test development (Hunt, 2011) Its use helps to reduce and refine the number of questions into a reduced set that tend to fit together into a tightly-related knowledge area (Cattell, 1978). Thus, the test becomes shorter and more parsimonious as it identifies and organizes the questions that ought to be retained from among those that were piloted. This method requires a large sample size for piloting. Although strict rules regarding sample size for exploratory factor analysis have largely disappeared (MacCallum, Widaman, Zhang, & Hong, 1999) sample sizes of at least 10 times the number of items analyzed is often deemed minimally acceptable (Child, 1990). Therefore, having a sufficient quantity of test results to perform exploratory factor analysis requires results from at least 180 participants to pilot an 18-item assessment. For clarity, we will next describe how to perform factor analysis with SPSS software.

After the data is opened in SPSS, click on "Analyze" from the top bar and select "Dimension Reduction." Then, select "Factor." Next, select the questions for analysis and click "Rotation." Click on "Rotation" and select "Varimax" because the data should be considered normalized. Also be certain to select "Rotated solution" so that this table will appear in the output. This method of rotation makes it easy to identify each variable with a single factor and is a commonly selected rotation option. Examine the output for the "Rotated Component Matrix." It will report the factor values for each component/factor with an Eigenvalue of 1.0 or greater. Eigenvalues report how well each variable loads on the factor and the amount of variance explained by each component (Forshaw, 2007). In the column of data for each component, the factor values for each variable listed in rows should be .50 or higher and should be at least .20 higher than its next highest column score. Otherwise, discard that variable from further consideration. To be useable, each component column must have at least two variables that survive this culling. Do not combine questions from different columns. Also, discard questions that cover concepts seemingly unrelated to other questions in the group to help assure content validity. In selecting which questions to eliminate, keep in mind that validity is a matter of degree rather an absolute. More than one set of questions might be useable and the creation of multiple measures, or constructs, is possible. (Nunnally, 1967)

For clarity, we display an example of factor analysis results in Table 1, titled Select Factor Analysis Results. The analysis displays data for 10 variables and 200 observations. Therein, we followed the steps listed in the two previous paragraphs. The first grid is titled "Total Variance Explained." Here, two components are indicated with Eigenvalues above 1.0. The next grid is titled "Rotated Component Matrix." The first component includes variables Q1 through Q8, all with values above .50. The second component includes variables Q9 and Q10. However, the component value for variable Q10 is below .50 and should be eliminated because it has a negative value. Component 2 is reduced to just one variable and should be discarded because single variable solutions should not be used.

TABLE 1
SELECT FACTOR ANALYSIS RESULTS

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.306	53.058	53.058	5.306	53.058	53.058	5.213	52.133	52.133
2	1.055	10.554	63.611	1.055	10.554	63.611	1.148	11.478	63.611
3	.893	8.928	72.539						
4	.753	7.532	80.071						
5	.706	7.057	87.128						
6	.357	3.570	90.698						
7	.345	3.446	94.144						
8	.284	2.835	96.979						
9	.187	1.868	98.847						
10	.115	1.153	100.000						

Extraction Method: Principal Component Analysis.

Rotated Component Matrix

	Component	
	1	2
Q1	.873	.066
Q2	.868	.092
Q3	.822	.146
Q4	.820	.030
Q5	.768	.098
Q6	.783	.028
Q7	.750	.084
Q8	.731	.130
Q9	-.215	-.631
Q10	-.064	.825

Reliability Analysis

Reliability of the groups of questions selected using exploratory factor analysis is determined through the calculation of Cronbach's Alpha scores. When the data is opened in SPSS, click on "Analyze," then "Scale," and finally, select "Reliability Analysis." Next, choose the items for each group of questions identified with exploratory factor analysis, then click on "Statistics." Then, click on "Scale if item deleted." Finally, click on "Continue" and finally, on "OK." Inspect the "Reliability Statistics" table for the initial Cronbach's Alpha. If it is above .50, the results are at least minimally acceptable (Haggerty & Denomme, 1991), but above .70 is preferable (Forshaw, 2007). If a higher Cronbach's Alpha is sought, inspect the table titled "Item-Total Statistics" for items, which if deleted, would result in a higher score. Repeat this step until the Cronbach's Alpha statistic is deemed acceptable or when fewer than two items result. In the event acceptable results are not produced, the entire process must be repeated, beginning with question selection and piloting.

We display an example of reliability analysis in Table 2, titled Select Reliability Results. The first grid is titled "Reliability Statistics" and displays a Cronbach's Alpha statistic of .923 for the eight items evaluated. This statistic exceeds the preferable threshold of .700. The second grid is titled "Item-Total Statistics." When one examines the last column in this grid, titled "Cronbach's Alpha if Item Deleted", all the reliability scores are smaller than the initial solution with all eight items. Therefore, the reliability cannot be improved and all eight items can be used as a reliable assessment instrument.

TABLE 2
SELECT RELIABILITY RESULTS

Reliability Statistics

Cronbach's Alpha	N of Items
.923	8

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Q1	26.28	7.999	.825	.906
Q2	26.32	7.855	.823	.906
Q3	26.35	7.966	.773	.910
Q4	26.33	7.730	.759	.912
Q5	26.33	8.041	.705	.916
Q6	26.24	8.425	.712	.915
Q7	26.26	8.585	.682	.917
Q8	26.25	8.422	.656	.919

CONCLUSIONS

In this paper, we suggest that a college of business might create its own assessment tests with proven reliability as an economical and effective alternative to the purchase of the ETS Major Field Test for Business. We described a method in keeping with accepted psychometric theories whereby multiple choice questions are selected and then piloted. Then, the resulting data are processed with exploratory factor analysis to determine possible useable questions that are then analyzed with reliability analysis to compute Cronbach's Alpha statistics.

REFERENCES

AACSB International. (2010). *Eligibility Procedures and Accreditation Standards for Business Accreditation,* Tampa, FL: Author.

Banta, T. A. & Associates (2002). *Building a Scholarship of Assessment*, San Francisco, CA: Jossey-Bass.

Banta, T. A. (2007). Warning on Measuring Learning Outcomes. *Inside Higher Education*, January 26: http://www.insidehighered.com/views/2007/01/26/banta Banta, T. A. (ed.) (2007). *Assessing Student Learning in the Disciplines*, San Francisco, CA: Jossey-Bass.

Banta, T. A., Jones, E. A. & Black, K. E. (2009). *Designing Effective Assessment: Principles and Profiles of Good Practice*, San Francisco, CA: Jossey-Bass.

Banta, T. A. & Blaich, C. (2011). Closing the assessment loop. *Change*, January/February, 22-27.

Banta, T. A. (ed.) (2011). *A Bird's-Eye View of Assessment: Selections from Editor Notes,* San Francisco, CA: Jossey-Bass.

Cangelosi, J. S. (1990). *Designing Tests for Evaluating Student Achievement,* White Plains, NY: Longman.

Cattell, R. (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences,* New York, NY: Plenum Press.
Child, D. (1990). *The Essentials of Factor Analysis:* London: Cassell Education Ltd.
Dwyer, C. A. (ed.) (2008). *The Future of Assessment, Shaping Teaching and Learning*, New York, NY: Taylor & Francis Group, LLC.
ETS. (2011a). *Major Field Tests*, Retrieved from http://www.ets.org/Media/Tests/MFT/pdf/mft_reliability_sem.pdf
ETS. (2011b). *About the ETS Major Field Tests*, Retrieved from http://www.ets.org/mft/about/
ETS. (2011c). *Pricing and Ordering*, Retrieved from http://www.ets.org/mft/pricing/
ETS. (2011d). *Find out how to prove –and improve – the effectiveness of your business program with the ETS Major Field Tests*, Retrieved from http://www.ets.org/Media/Tests/MFT/pdf/mft_testdesc_business_4cmf.pdf
Ewell, P. T. (1984). *The self-regarding institution: Information for excellence,* Boulder, CO: National Center for Higher Education Management Systems.
Forshaw, M. (2007). *Easy Statistics in Psychology*, Malden MA: Blackwell Publishing.
Haggerty, T. R. & Denomme, D. (1991). Organizational Commitment in Sport Clubs: A Multivariate Exploratory Study. *Journal of Sport Management*, 5 (1), 58-71.
Hoover, J. D., Giambatista, R. C., Sorenson, R. L. & Bommer, W. H. (2010). Assessing the Effectiveness of Whole Person Learning Pedagogy in Skill Acquisition. *Academy of Management Learning & Education*, 9, 192-203.
Hunt, E., (2011) *Human Intelligence*, New York, NY: Cambridge University Press.
Lyman, H. B. (1986). *Test Scores and What They Mean (4th)*. New York, NY: Prentice-Hall.
MacCallum, R. C., Widaman, K. F., Zhang, S. & Hong S. (1999). Sample size in factor analysis. *Psychological Methods*, 4(1), 84-99.
McMillan, H. H. (2008). *Assessment Essentials for Standards-Based Education*, Thousand Oaks, CA: Corwin Press.
Messick, S. (1993). Trait Equivalence as Construction Validity Across Multiple Methods of Measurement. In R.E. Bennett & W. C. Ward (Eds.) *Construction versus choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment,* Hillsdale, NJ: Lawrence Erlbaum Associates.
Nunnally, J. C. (1967) *Psychometric Theory*, New York, NY: McGraw Hill.
Pascarella, E. T. & Terenzini, P. T. (1991). *How College Affects Students,* San Francisco, CA: Jossey-Bass.
Perkins, D. (1999). The many faces of constructivism. *Educational Leadership*, *57*(3), 6-11.
Pike, G. R. (2006). The Convergent and Discriminant Validity of NSSE Scalelet Scores. *Journal of College Student Development,* 47, 550-563.
Raykov, T. & Marcoulides, G. A. (2010) *Introduction to Psychometric Theory,* New York, NY: Routledge.
Tuckman, B. W. (1985). *Evaluating Instructional Programs (2nd)*, Newton, MA: Allyn and Bacon, Inc.
Walvoord, B. E. (2010). *Assessment Clear and Simple: A Practical Guide for Institutions, Departments, and General Education (2nd),* San Francisco, CA: Jossey-Bass.
Willingham, W. W., Young, J. W. & Morris, M.M. (1985). *Success in College: The Role of Personal Qualities and Academic Ability*, New York, NY: The College Entrance Examination Board.