Research Report

Repurposing a Business Learning Outcomes Assessment to College Students Outside of the United States: Validity and Reliability Evidence

Guangming Ling

December 2013

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.

Repurposing a Business Learning Outcomes Assessment to College Students Outside of the United States: Validity and Reliability Evidence

Guangming Ling

Educational Testing Service, Princeton, New Jersey

December 2013

Action Editor: Donald E. Power

Reviewers: David Klieger and Jonathan Steinberg

Abstract

One concern when repurposing a test to a new population is whether the test is measuring the same construct in a valid and reliable way that is comparable to the intended population. Following the guidelines of the International Test Commission and the ETS Standards for Quality and Fairness, this study was designed to collect evidence in support of repurposing the Major Field Test for the Bachelor's Degree in Business (MFT-B) for use in undergraduate level business programs outside of the United States. The author examined the test's reliability and internal structure based on 930 senior college students from 9 business programs in Asia, Europe, the Middle East, and South America. Each participating institution satisfied 2 requirements for this study: (a) having a bachelor's level business curriculum that is similar to their U.S. peer programs and (b) using English as the instructional language. The analyses of the data showed that the MFT-B has the same factor structure for the non-U.S.–based sample as that for their U.S. peers. In addition, the test scores for the non-U.S.–based sample are highly reliable (.90) overall, with some variations among the 9 programs. Both types of evidence support the notion that the MFT-B is appropriate when repurposed for use with non-U.S.–based business programs.

Key words: test repurposing, non-U.S. –based business programs, learning outcomes test, reliability, construct validity

i

In recent years, learning outcomes in higher education have gained attention from education providers, policy makers, and public audiences in the United States and across the world (United States Department of Education, 2006). Other than comparing students' learning outcomes across classes or cohorts within a particular program, there are also interests in making comparisons with peer programs on the learning outcomes, particularly with similar programs in the same or different countries. This poses an important question about the comparability of learning outcomes of programs across countries and regions. A way to answer this question is to develop a common measure to facilitate the comparison of similar programs across countries. A good example is the Program for International Students Assessments, or PISA (Organisation for Economic Co-operation and Development, 2010), which is currently administered in 65 nations to facilitate country-level comparisons on pupils' scholastic performance. However, it is difficult and costly to develop such a common measure and to account for the vast diversity among programs and countries, which requires a high degree of localization and customization related to language, culture, and curriculum design. A practical and reasonable compromise might be to consider repurposing a test that is already established in one country to other countries or regions, where these challenges may be relatively easier to manage.

Repurposing a test to an unintended population or for an unintended use may require a series of investigations and adequate evidence to ensure test validity and fairness. Wendler and Powers (2009) suggested that "using a test either for test takers or for purposes that are different from those for which the test was originally developed" (p. 1) is considered repurposing the test. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) require that test scores be investigated on their fairness and consequences prior to being used for any purposes in unintended populations (Standards 1.4 and 1.5). The ETS Standards for Quality and Fairness also emphasize the importance of collecting the necessary evidence when test-taker populations change (Educational Testing Service [ETS], 2002, Standards 6.7; ETS, 2007). Similar requirements can be found in the adaptation guidelines suggested by the International Test Commission (2000).

Although repurposing a test can eliminate the time and expense of developing a new common measure, the repurposing approach cannot eliminate the diversity-related issues found across countries, nor can it avoid variations in culture, language, and curriculum design. The goal

of this study is to address some of these issues by examining the construct validity and reliability of a U.S.-based business learning outcomes assessment (the *ETS*® Major Field Test in Business [MFT-B]; ETS, 2010) if it is repurposed to business students outside of the United States. For convenience purposes, the term *international business students* is used hereafter to represent business students who study in business programs in a non-U.S. location, whether or not the business program is affiliated with a U.S.-based college or university.

The MFT-B is a learning outcomes measure intended for business students upon completion of their major-related courses (ETS, 2010). The test consists of business-related content areas, such as accounting, economics, management, quantitative business analysis and information systems, finance, marketing, and legal and environmental issues, as reflected by the national business curriculum survey results based on hundreds of U.S. programs. Individual students' total scaled scores and the group level assessment indicator scores (AIs; ETS, 2010; Ling, 2012) are reported. The MFT-B is a highly regarded learning outcomes assessment valued by more than 600 business programs in the United States, most of which have been accredited by the Association to Advance Collegiate Schools of Business (AACSB International) or other accreditation agencies.

Five requirements (Lord, 1980) are often used in the educational testing field to examine the comparability of scores of a repurposed test (see also Holland & Dorans, 2006; Kolen & Brennan, 2004). First, the test must measure the same construct in the two populations (the intended and repurposed populations). This provides construct-related evidence to support claims that the test measures the same ability/skill, and that the internal test structure is the same in the two populations (see also Wendler & Powers, 2009). The second requirement is that the test scores based on the two populations have the same level of reliability. The remaining three requirements are related to the equating (or linking) of test scores based on multiple test forms. The equating transformation within the intended or repurposed sample should have a symmetric property so that the equating function works in the same way in reverse (the Symmetry Requirement; Lord, 1980, p. 199). The equating function between the two forms should work in the same way in the intended and repurposed population (the Population Invariance Requirement; Lord, 1980, p. 199). As for the equating results, it should not matter which form a given student takes or to which population the test taker belongs (the Equity Requirement; Lord, 1980, p. 199).

From the perspective of test use, a test (i.e., one test form) satisfying the first two requirements as defined by Lord (1980) can support the appropriateness of using the test in the intended population (e.g., the U.S.-based business students taking the MFT-B) and the repurposed population (e.g., the non-U.S.–based business students), which would further support the appropriateness of comparing test scores between these two populations or any subsamples of each population and make inferences related to business-related learning outcomes targeted by the test. This will have practical implications for the use of MFT-B scores, as the U.S. business programs often compare their MFT-B scores among themselves based on the Comparative Data Reports (ETS, 2013). The goal of this study is to examine whether the MFT-B administered to the non-U.S.–based business students has satisfactory psychometric properties according to Lord's requirements, which could help to decide whether it is appropriate to compare the MFT-B scores between the international business programs and the U.S.-based business programs.

More specifically, the study aimed to answer the following questions:

1. Do scores from the MFT-B have a high degree of reliability for international business students?
2. Do scores from the MFT-B have the same internal structure for international business students as for U.S. business students?

Method

Sample

Nine international business programs from around the world were selected based on two criteria: (a) English is used as the primary language for instruction and coursework and (b) the business curriculum is similar to those in the U.S.-based business programs and covers all of the content areas targeted in the MFT-B. These criteria ensured that students would be affected by taking the test in English to a limited degree, and that all content and knowledge areas in the test are covered by the participating programs.

Among the programs selected, four were from the Middle East, one from South America, one from Europe, and three from Asia. Eight of the programs were on satellite campuses of American universities or were joint programs with a business program located in North America, which ensured the same curriculum designs and teaching methods as the North American business programs. Seven of the schools had been accredited by the AACSB International prior

3

to this study, which means these programs have met the accreditation requirements and are comparable to their U.S. peer programs in terms of the educational quality (AACSB, 2013).

A total of 930 students participated in this study, with the sample sizes of the programs varying from 25 to 169. Among all the participants, 39% were males; more than half (69%) said they communicate better in English or communicate equally well in English and in another language. Among the nine programs, the percentage of males varied from 21% to 56%, and the percentage of those who reported communicating better in English or equally well in English and another language varied from 36% to 87%. (See Table 1.)

Table 1

List of Programs and Descriptive Statistics

Region	Middle East				South America	Asia			Europe	Non-U.S.	U.S.
Program	A	B	C	D	E	F	G	H	I		
N	169	89	172	150	51	25	134	36	104	930	15,523
Male	40%	47%	30%	47%	56%	42%	37%	21%	49%	39%	47%
English	87%	88%	61%	76%	47%	50%	43%	36%	64%	69%	90%
M	66.07	42.93	50.01	62.79	65.65	40.08	44.76	63.17	51.44	54.82	56.92
SD	12.72	11.38	12.19	14.81	9.69	11.13	15.81	13.36	11.32	15.84	14.98
Reliability	0.87	0.83	0.85	0.90	0.78	0.84	0.91	0.89	0.82	0.91	0.89

Instruments

A screening survey was used to collect information from the overseas business programs, such as the primary instruction language and the curriculum design. An online version of the MFT-B was administered to the students in each program, followed by a student background information questionnaire (BIQ) that is also used in the United States. The MFT-B test includes 120 multiple-choice items in total. These items cover seven content areas of business knowledge and skills: (a) accounting, (b) economics, (c) management, (d) quantitative business analysis and information systems, (e) finance, (f) marketing, and (g) legal and social environment, labeled as S1 to S7, respectively. The number of items varies from 12 to 21 across the seven content areas (see Ling, 2012, for details).

Data and Analysis

Descriptive analyses were conducted to examine the distributions, means, and standard deviations of the total number of correct scores. Reliability analyses of the total test score were also conducted for all students and subgroups indicated by best language status and business programs. As two items were excluded from the U.S.-based operational score reporting because they were either too easy or too difficult, the reliability and internal structure analyses in this study were conducted with the remaining 118 items. Score differences associated with background variables indicated by the BIQ questions, including gender and best language, were analyzed descriptively firstly and then using a general linear model (GLM) approach. Cohen's effect size (d) was computed to examine whether the mean difference was of any practical importance. Cohen's d is considered small for values between .2 and .5, moderate for values between .5 and .8, and large or substantial beyond .80 (Cohen, 1988).

Structural equation modeling (SEM) methods were used to examine the internal structure of the test when taken by the international students. Similar to Ling (2012), a measurement model was constructed separately for each of the seven content areas, where all the items associated with a content area were set as indicators of a latent factor. For example, there were 21 items on accounting-related business knowledge and skills, which were all set to load on a common latent variable of accounting-related business knowledge (see Ling, 2012, for details). In addition, a single-factor model confirmed by Ling using a U.S. sample was also fitted to the data, where all of the 118 items were set to load on the same common latent variable representing business-related knowledge. The single-factor model was expected to fit the data well in support of the claim that the test has the same internal structure for both the U.S. and non-U.S. student population.

Three fit indices were used to evaluate the model-data fit: root mean square error of approximation (RMSEA), comparative fit index (CFI), and the Tucker–Lewis index (TLI). Some empirical guidelines were followed when evaluating these fit indices: a model with an RMSEA value below .08, a TLI value above .90, a CFI value above .90, were considered to be an acceptable fit; a model with an RMSEA value below .05 and a TLI (and CFI) value above .95 can be considered a good fit (Browne & Cudeck, 1993; Hooper, Coughlan, & Mullen, 2008; Hu & Bentler, 1999; Raykov & Marcoulidies, 2000; Yu & Muthén, 2001).

Results

The MFT-B demonstrated high reliability (internal consistency) when used for business programs outside of the United States, with a Cronbach's alpha coefficient value of .91 based on the 118 items. For each program, Cronbach's alpha values were all above .83, except for one program (α = .78, see Table 1). The test's reliability based on students grouped by their best language was also high, .91 for students who said they communicated better in English, .90 for those who communicate better in another language, and .92 for those who communicate equally well in English and another language.

Overall, the international students answered slightly less than half of all items correctly (M = 54.82, see Table 1), slightly lower than that of their U.S. peers (56.92). Such a difference is .14 (Cohen's *d*) if put in the unit of *SD* of the U.S. students, which is rather trivial according to Cohen (1988). Among the nine programs, the average scores ranged from 40.08 (SD = 11.13) to 66.07 (*SD* = 12.72, see Table 1); the differences among the nine programs were significant—*F* (8,921) = 55.81, *p* < .001. Further post hoc multiple comparisons using the Bonferroni procedure and the Type I error of .05 suggest that the mean scores of most (27) pairs of programs were significantly different, indicating a diversity on the learning outcomes among programs included here.

Performance of Subgroups

Further comparisons were made for schools grouped by region. This is mainly to explore, assuming the programs of this study are representative of similar programs in each region, whether there exists any differences among the regions on the total scores. Significant differences on the mean scores were found among the four regions: *F* (3, 926) = 28.92, *p* < .001. The follow-up post hoc multiple comparisons among the regions were performed using the Bonferroni procedure with a Type I error of .05 (or .008 for each comparison among the six pairs). More specifically, South American students scored the highest on average (65.65, *SD* = 9.69), significantly lower than the other regions. The students from the Middle East scored 56.91 on average (*SD* = 15.64), significantly greater than that of students from Europe and Asia; and the students in Europe (*M* = 51.44, *SD* = 11.32) and Asia (*M* = 47.56, *SD* = 16.64) scored with no significant difference.

Overall, there was no significant difference between male students and female students (56.20 vs. 55.53), $F (1, 764) = 5.94$, $p = .871$, $d = .06$ on the mean total scores, excluding those with missing gender information. There was a significant interaction effect between gender and region, $F (3, 958) = 3.85$, $p = .008$ on the total score. This significant interaction appears to be due to the fact that, in Asia, females performed significantly better than males (51.31 vs. 44.52, $d = .41$; $F [1, 173] = 6.41$, $p < .012$), whereas there were no significant gender differences on the total score in the other regions (see Figure 1).
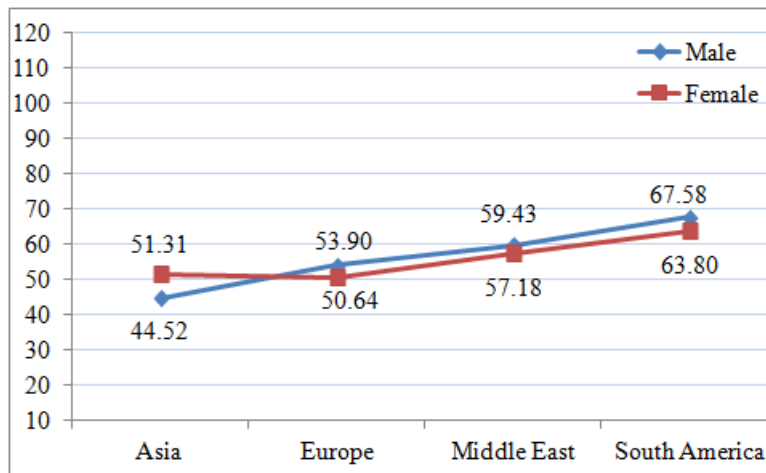

Figure 1. **Gender difference on the total MFT-B score by region.**

Overall, significant differences on the MFT-B scores were found to be associated with the self-reported best-language status, $F (2, 927) = 17.97$, $p < .001$. Further post hoc multiple comparisons using the Bonferroni procedure ($p = .05$) showed that students who reported communicating equally well in English and another language scored significantly higher than those who reported that they communicate better in English or those reporting they were better in another language, but the latter two groups were not significantly different on the total score. However, after adding the region as a fixed factor, the best-language–related effects on the total scores were no longer significant, $F (2, 927) = .780$, $p = .474$, nor was the interaction between best language and region, $F (6, 923) = 1.77$, $p = .104$. That is, the nonsignificant best-language–related main effects on the MFT-B scores were not significantly different among the regions.

Results Based on Structural Equation Modeling

As was mentioned earlier, a measurement model was fitted to the data of each of the seven content areas, similar to Ling (2012). The measurement model fit the data acceptably well for six content areas, with the other area having marginally acceptable values on the fit indices (see Table 2). The RMSEA was in the range of .018 to .034, the TLI in the range of .91 to .98 (except for the subscale for the legal and social environment area, TLI = .89, see Table 2). The CFI was in the range of .94 to .98 except for the quantitative business analysis and information system (.89) and the legal and social environment area (.87).

A single-factor model with all 118 items loaded on the same latent variable of business-related knowledge and skill was fitted to the data, with acceptably good fit indices. The RMSEA value was .015, and the TLI and CFI values were .94 and .94, respectively (see Table 2), which are similar to those based on the U.S. sample as found in Ling (2012).

Table 2

Fit Indices for SEM Models for the Test and Each Content Area

Subscale	N items	RMSEA	90% CI RMSEA	TLI	CFI
S1-Accounting	21	.023	.017–.029	.97	.96
S2-Economics	20	.018	.008–.025	.98	.98
S3-Management	19	.023	.017–.029	.95	.94
S4-Quantitative Analysis & Information Systems	19	.029	.019–.039	.91	.89
S5-Finance	13	.028	.021–.034	.96	.95
S6-Marketing	14	.024	.017–.031	.96	.95
S7-Legal and Social Environment	12	.034	.025–.042	.89	.87
Total test: 1-factor model (Non-U.S.)	118	.015	.014–.016	.94	.94
Total test: 1-factor model (U.S.)	118	.015	.014–.016	.96	.90

To summarize the results, the MFT-B test had high reliability when used with international business students. Slight variations in the scale reliability coefficients were found among the nine programs. The average scores for the international students were slightly lower than their U.S. peers. The mean scores differed among programs and regions. Overall, no gender differences were found. However, female students scored significantly higher than male students in Asia. Students who reported communicating equally well in English and another language performed significantly better than those who communicate better in another language or those who communicate better in English, although these differences were no longer significant after controlling for region. Overall, a single-factor measurement model that was previously confirmed with the U.S.-based students was also confirmed with the international students.

Discussion

In general, the findings of this study confirmed that the MFT-B measures the same unidimensional structure of business-related knowledge and skills for non-U.S.–based business students as their U.S. peers; the test scores of the international students had the same high reliability as those of the U.S. business students. Both types of evidence support the notion that the MFT-B is appropriate when repurposed for use with international business students.

It seems that the criteria used to screen non-U.S.–based business programs helped to identify a group of international business programs for which the use of the MFT-B is appropriate. The use of English as an instructional language appears to be an important condition for non-U.S.–based business programs, as it can screen out students with English proficiency levels too low to answer the MFT-B items in English. However, the non-U.S.–based sample had a relatively lower mean score than the U.S. test-taker population on the MFT-B, although the difference appears to be practically trivial. One possible explanation is that some students included in this study have a degree of English proficiency that is not as developed as for their U.S. peers, which may have affected, to some extent, their test performance. Another plausible explanation is that the content covered or focused on in non-U.S.–based business programs may slightly differ from that in the United States, mainly because the business context in these countries and regions differ from the U.S. context to varying degrees.

9

An interesting finding was that, overall, those students who reported communicating equally well in English and another language scored higher than those who communicate better in another language or those who communicate better in English, but no significant difference was found between the latter two groups of students. It is not clear whether this only applies to non-U.S.–based business students or whether there is any other factor in these business programs not considered in this study but which may help explain such difference. The fact that such differences disappeared after taking into account region-related variations suggests factors other than English language proficiency may also be related to the learning outcomes as reflected on the MFT-B scores. For example, students who communicate equally well in English and another language may be more capable as a result of the self-selection, and it is possible they are more motivated and better performing students academically, which could explain why their scores on the MFT-B were much higher than those who communicate in English.

Regional differences or program level differences may have some implications in interpreting the test scores in the context of an international setting. Such variations may depict a true picture of individual differences among programs, as seen among U.S. business programs. However, possible differences in English language proficiency levels and in curriculum design may result in greater variation for non-U.S.–based programs than for the U.S. programs. The extent to which the program level variations of non-U.S.–based programs resemble those of the U.S. programs remains a question. A more definitive answer to such a question may be obtained when data from more non-U.S.–based business programs (i.e., the numbers of students and programs) are available.

Finally, females were found to have a moderately greater mean score than males among business students studying in Asia, but not in other regions. Such a trend is different from other studies based on students studying in the U.S.-based business programs. For example, Bielinska-Kwapisz & Brown (2012) found that male business students performed better than female business students on the total MFT-B score. It is not clear about the factors that may be related to this finding among students in Asia.

Several limitations of the current study should be noted. First, although the sample sizes of students may be adequate for the statistical analyses, the number of programs may be increased to make the results more generalizable. An alternate option might be to continue monitoring the psychometric properties (reliability and internal structure) with more schools

using the test in an operational setting. Second, the institutions or business programs included in this study were all self-selected, but satisfied the two selection criteria, which may be more representative if more programs from South America, Europe, and the other regions can be included, especially on the region-related differences. Finally, the current study only examined the internal structure and the reliability of a single form of the MFT-B. That is, it did not examine comparability issues associated with multiple forms, which may need to be addressed with testing data from a larger sample of non-U.S.–based students on multiple forms of the MFT-B.

Despite these limitations, two tentative conclusions may be reached, given the findings of this study. First, though institutional or regional differences exist on the test scores, the findings of this study support the use of MFT-B for international business students, as the test reliability was high, and the internal structure of the test was the same as that based on the U.S. peer students. In other words, the current findings support the use of MFT-B test scores to compare the non-U.S.–based business students with the U.S. students (e.g., using comparative data) and make inferences on business-related learning outcomes. Second, the two requirements regarding the instructional language and the business curriculum should also be applied for international business programs that are interested in using the MFT-B.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Association to Advance Collegiate Schools of Business. (2013). *2013 accreditation standards*. Retrieved from: http://www.aacsb.edu/accreditation/2013standards/

Bielinska-Kwapisz, A., & Brown, F. W. (2012). Differential gender performance on the Major Field Test—Business. *Journal of Education for Business*, *88,*159–166.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

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

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Educational Testing Service. (2007). *ETS international principles for fairness review of assessments.* Princeton, NJ: Author.

Educational Testing Service. (2010). *About the Major Field Tests: Test content overview*. Retrieved from http://www.ets.org/mft/about/content

Educational Testing Service. (2013). *Comparative data report*. Retrieved from http://www.ets.org/mft/scores/compare_data/

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: American Council on Education.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*, 53–60.

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

International Test Commission. (2000). *TC test adaptation guidelines.* Retrieved from http://www.interstcom.org/guidlines/test+adatation.php

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Ling, G. (2012). *Why the Major Field Test in Business does not report subscores of individual test takers: Reliability and construct validity evidence* (Research Report No. RR-12-11). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Organisation for Economic Co-operation and Development. (2010). *PISA 2006 technical report.* Retrieved from http://www.oecd.org/dataoecd/0/47/42025182.pdf

Raykov, T., & Marcoulidies, G. A. (2000). *A first course in structural equation modeling.* Mahwah, NJ: Erlbaum.

United States Department of Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. Washington DC: Author.

Wendler, C., & Powers, D. (2009). *What does it mean to repurpose a test?* (ETS R&D Connections. No. 9). Princeton, NJ: Educational Testing Service.

Yu, C. Y., & Muthén, B. O. (2001). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes* (Technical report). Los Angeles, CA: University of California at Los Angeles, Graduate School of Education and Information Studies.