6th International Conference on Higher Education Advances (HEAd'20)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/HEAd20.2020.11303

Reliability of multiple-choice versus problem-solving student exam scores in higher education: Empirical tests

Eric Lee[1], Naina Garg[2]
[1]Department of Finance, Information Systems, and Management Science, Saint Mary's University, Canada, [2]Masters of Financial Economics graduate, University of Toronto, Canada

Abstract

Instructors in higher education frequently employ examinations composed of problem-solving questions to assess student knowledge and learning. But are student scores on these tests reliable? Surprisingly few have researched this question empirically, arguably because of perceived limitations in traditional research methods. Furthermore, many believe multiple choice exams to be a more objective, reliable form of testing students than any other type. We question this wide-spread belief. In a series of empirical studies in 8 classes (401 students) in a finance course, we used a methodology based on three key elements to examine these questions: A true experimental design, more appropriate estimation of exam score reliability, and reliability confidence intervals. Internal consistency reliabilities of problem-solving test scores were consistently high (all > .87, median = .90) across different classes, students, examiners, and exams. In contrast, multiple-choice test scores were less reliable (all < .69). Recommendations are presented for improving the construction of exams in higher education.

Keywords: *Exams; internal consistency; reliability, higher education.*

1. Introduction

Many instructors believe multiple-choice (MC) tests offer advantages over other question types (e.g., problem solving or essay): minimal effort to grade, reliability, and validity. We question this wide-spread belief. Many instructors in diverse fields construct what, in our opinion, are poor MC questions (e.g., poorly written alternatives and stems). We suspect that to use MC tests effectively, one must be trained psychometrically. In contrast, instructors in many disciplines commonly employ exams composed of problem-solving (PS) questions (Garg & Lee, 2016). We believe that such exams may provide an attractive alternative for many courses (though they certainly take longer to mark). PS questions offer advantages: they encourage higher-order critical thinking, typically take little time to construct, and seem well suited to the teaching demands of smaller, advanced courses or quantitative courses. We believe student scores on such exams are highly reliable, but are they? The purpose of this paper is to empirically assess the reliability of student scores on PS subtests and compare it with that for MC subtests on the same exam. In a series of empirical studies in 8 classes (401 students) in business finance, we used a methodology based on three key elements to examine these questions: A true experimental design (Lee & Whalen, 2007), appropriate estimation of exam score reliability, and reliability confidence intervals. We first review literature on reliability of exams and then describe the empirical approach used to address our major research question, and end by discussing the results and conclusions.

2. Literature Review

Important academic decisions are based, at least in part, on the marks that students attain on their examinations. Exam scores must be reliable (Brennan, 2001). The greater the reliability of student exam scores, the more confident we can be that differences in grades assigned reflect actual differences in the knowledge and skills being assessed and not the result of random error (Dracup, 1997). The error associated with student scores on an exam increases as reliability decreases. As well, tests with more questions are generally more reliable (Ebel, 1972). "Reliability, however, is not a property of the test itself, but a property of a set of test scores, (p. 25)" (Frisbie, 1988). Thus, an exam is not characterized by a single reliability but varies with the set of test scores. Classical test theory deals with the reliability of classroom exam scores (Nunnally & Bernstein, 1994). Morley (2014) asserts that internal consistency reliability, derived from classical theory, "is appropriate when we want to make statements about the respondent" (students, in our case). It is a measure of how well exam questions assess the learning in an academic course. It can be assessed from a single administration of a test, and it is a frequently reported measure of reliability (Hogan et al., 2000).

Following common psychometric practice, we estimated internal consistency reliability using coefficient alpha (α). Alpha is appropriate for estimating reliability provided all questions on a test are of the same kind and all questions are equally weighted (assumptions required by the tau-equivalent measurement model), as is typically true for MC tests. However, when these requirements are not met, alpha can underestimate reliability (Dunn et al., 2014; Miller, 1995). Estimates derived from a congeneric measurement model are more appropriate when questions are of different types or vary in score value, as they do for mixed-format or PS tests. We, therefore, also used the most commonly recommended measure of congeneric reliability, coefficient omega ω, in these cases (Padilla & Divers, 2013).

One should set a standard for reliability (Fan & Thompson, 2001). We set a minimum target in our study at .70 given practitioners for high stakes MC professional tests use this criterion (Williams et al., 2004) and classroom exams can be for high stakes (from a student perspective). What research has been conducted on reliability of MC and PS exams? Cox (1967) observed that despite the importance of examining, "very few of those actively engaged in it regard it as a field for experiment and research". We found only one paper on PS exams. Hill (1978) assessed 3.0-hr PS final exam scores in engineering courses and concluded that reliability was very poor. However, Hill assessed inter-marker reliability, not internal consistency. His exams consisted of 5-6 questions. At least 10 separately markable questions should be posed on an exam if reliability is to exceed .70 (Garg & Lee, 2016).

We confine our discussion of studies examining MC reliability to four well-conducted ones. First, Jensen et al. (2013) found the reliability (α) of a final exam in introductory biology in two classes (155 students) to be .66, a reliability many would consider too low. Second, Royal & Hedgpeth (2015), for a MC medical exam, also found reliability to be low at .60 (α). Third, for exams in intro physics, Harrison (2014) reports that over the years they were never able to achieve a reliability > .70 for their MC tests. Finally, DiBattista and Kurzawa (2011) report reliabilities in a comprehensive study of MC exams in 16 courses. Half of the classes had reliabilities lower than .70. These results suggest that MC test score reliabilities may not, in practice, be as high as believed. But can PS test score reliability be higher?

3. Method

3.1. Courses, Classes, Students, Instructors, Exams

Over a period of 4 years, 401 students attended one of 8 classes over 2-4 months with 39 lecture-hours at a mid-sized Canadian university in Introductory Finance I (a second-year course). About 50% were male, 50% female, with most aged 19 to 25. International students comprised approximately 50% of all students. While most were in their second

year of an undergraduate business degree, some were in their third and fourth years. This course covered: financial analysis, capital budgeting, working capital management, the tax environment, and the role of financial intermediaries. Two instructors, one male and one female, taught these classes. Students wrote a 3.0-hour final exam, typically worth about 50% of the total marks for the course. Each of these four mixed-format exams consisted of different question types: MC, PS, and, in some cases, a few short-answer (SA) or true-false (TF) questions. The number of alternative answers for each MC question varied between 4 to 6, and each MC question was always worth the same on a given exam (either 1% or 2%). An example MC exam question follows: Con Artists, Inc. has just paid a dividend of $0.55 per share. The dividends are expected to grow at an annual rate of 5 percent indefinitely. How much should the stock be sold for today if the required return is 12.5 percent? (a) $4.62, (b) $7.23, (c) $7.70, (d) $11.55, (e) none of the above. PS questions varied markedly in marks awarded, ranging from 2% to 15%. Such questions were composed of 1 to 6 different parts (with each part worth between 1% and 10%). An example of a PS question follows: Simkins Inc has just developed a solar panel capable of generating 200% more electricity than any solar panel currently on the market. As a result, Simkins is expected to experience a 15% annual growth rate for the next 3 years. By the end of 3 years, other firms will have developed comparable technology, and Simkins growth rate will slow to 5% per year forever. Shareholders require a return of 12% on Simkins stock. The most recent annual dividend, which was paid yesterday, was $1.75 per share. (a) Calculate the current market value of Simkins stock. (b) Calculate the expected market price in one year. (c) Calculate the expected dividend yield and capital gains yield expected during the first year. The university's Research Ethics Board deemed the research acceptable.

3.2. Procedure

Data constituted 8 empirical data sets, one for each class. A data set consisted of a set of n student vectors in a class, each vector composed of marks awarded on each part of each separately markable question on an exam. Marks for each exam (or subtest) that did not sum to 100 marks were then, for comparability, renormalized to a range of 0 to 100%. Thus, a mark of 55 out of a maximum possible 110 on such an exam would result in a grade of 50%.

3.3. Reliability Assessment

To compute coefficients alpha and omega, we used the MBESS package (Dunn et al., 2014) in R. The normal theory bootstrap approach was used as it is superior to other estimates for the small sample sizes typical of our classes (Padilla & Divers, 2013). For questions on the MC subtests, reliability was estimated using coefficient alpha (α_{mc}). The appropriate measure of reliability for mixed-format and PS tests is omega. However, coefficient alpha was also assessed to explore the degree of underestimation of reliability by coefficient

alpha in mixed-format and PS exams. Given the few TF or SA questions, we did not investigate them.

4. Results

4.1. Correlational Analyses

The correlation between MC and PS subtest scores reflects the extent to which these tests measure the same construct (knowledge in these finance classes). Student performance on MC subtests, in general, correlated only moderately with that for PS subtests (median raw r = .40). Estimates of true-score correlations, computed by dividing raw score correlation by the square root of the product of the MC and PS subtest reliabilities (Nunnally & Bernstein, 1994; Charter & Feldt, 2002), were higher with a median of .70. Both sets of correlations suggest that MC and PS subtests were measuring somewhat different learning about finance.

4.2. Reliability Analyses

Table 1 displays the internal consistency reliability estimates for all final exam and subtest scores for the four exams in the 8 classes. As expected, for the full-length mixed-format exams, coefficient alpha underestimated reliability in all 8 classes (all ω_{ex}'s > α_{ex}, median difference = .02). As well, for PS subtests, alpha underestimated reliability in all 8 classes (all ω_{ps}'s ≥ α_{ps}, median difference = .025). Consequently, we relied on coefficient omega to assess reliability for PS and mixed-format tests. For mixed-format final exams, all point estimates of reliability (ω_{ex}) exceeded the high standard of .80 set as our target (all $\omega_{ex} \geq$ .87, median ω_{ex} = .89). The same pattern was repeated for student scores on PS subtests (all $\omega_{ps} \geq$ .87, median ω_{ps} = .90). In contrast, for MC subtest scores, none of the point estimates of reliabilities (α_{mc}) met even the minimal target of .70 (all α_{mc} < .70, median α_{mc} = .41).

4.3. Spearman-Brown reliability analyses

For the exams administered, each MC subtest consisted of fewer questions, fewer total marks allocated, and less time allotted to answer than the PS subtest on the same examination. A fairer comparison between MC and PS subtest reliabilities would be to estimate the reliabilities of MC subtests equated on subtest length. Therefore, using the Spearman-Brown prophecy formula (Nunnally & Bernstein, 1994), we predicted the reliability (coefficient alpha) of a suitably lengthened MC subtest equated for the same final exam on the time allotted to complete each subtest or, equivalently, on the number of marks allotted since time and marks allotted were always confounded in our studies. The median Spearman-Brown reliability point estimate was .72 (mean = .70), suggesting that MC test scores, at least as constructed by these examiners, were typically not highly reliable.

Table 1. Reliabilities (α and ω) for student marks on MC subtests, PS subtests, and final exams

Class	Ex	n	α_{ex}	ω_{ex}	α_{mc}	α_{ps}	ω_{ps}
F1	1	48	.89	.91	.69	.86	.90
F2	1	49	.84	.87	.66	.85	.88
F3	2	48	.88	.89	.06	.88	.89
F4	2	51	.86	.88	.19	.85	.87
F5	3	43	.88	.89	.41	.89	.90
F6	3	56	.90	.91	.52	.90	.90
F7	4	48	.85	.89	.41	.85	.89
F8	4	58	.88	.92	.46	.88	.92

Note. Ex = exam; *n* = number of students; ω_{ex} = coefficient omega for final exam composed of all types of questions; α_{mc} = alpha for mc subtest; α_{ps} = alpha for ps subtest; ω_{ps} = omega for PS subtest.

4.4. Differences Between Reliabilities Analyses

For each class separately, we tested the difference between unadjusted reliabilities for MC and PS subtests using Feldt's (1980) repeated-measures *t*-test (null hypothesis H_0: $\alpha_{ps} = \alpha_{mc}$). The mean coefficient alpha for PS subtests were significantly higher (.87) than that for MC subtests (.70) in all 8 classes (all *p*'s < .05). Equating MC and PS subtests for time allotted to complete the subtest (or equivalently, worth the same number of marks), we found reliability for PS subtests to be significantly higher than that for MC subtests in 6 of the 8 classes (all except classes 1 and 2), using an *F*-test (H_0: $\alpha_{ps} = \alpha_{mc/SBtime}$) of the difference between dependent alpha reliabilities (adjusted using the Spearman-Brown formula to equate on number of test questions or test time) (Alsawalmeh & Feldt, 2000). α_{ps} underestimates the true reliability of the PS subtests since coefficient omega always equals or exceeds coefficient alpha for congeneric data (Hogan et al., 2012). Consequently, all of these statistical tests underestimate the significance of the difference in reliabilities between PS and MC subtests.

5. Discussion

Given the importance of making good decisions affecting the future of students, we ask how reliable our exams really are and whether we can increase the internal consistency reliability of grades awarded. Coefficient alpha, the most commonly used measure of reliability, often underestimates it. Instead, reliability of problem-solving exams is best estimated using a congeneric measurement model value such as coefficient omega (Qualls, 1995).

As predicted, PS subtest scores were consistently highly reliable across different classes, students, instructors, and exams (all $\omega_{ps} \geq .87$, median $\omega_{ps} = .90$). Furthermore, reliabilities for PS subtests (ω_{ps}) were always higher than those for MC subtests (equated on time for students to complete subtest, $\alpha_{mcSBtime}$) on the same final exam in all 8 classes. Why are PS tests consistently more reliable? We suspect that MC exams require extensive psychometric training (DiBattista & Kurzawa, 2011) whereas PS exams do not. Are MC tests the best way forward? Many believe so. Our review of the literature and our results argue against this position. MC score reliabilities were often below .70. While MC reliabilities can be improved with training, PS tests require no training. MC and PS subtests were only moderately correlated with one another. They seem to be assessing different knowledge on typical finance examinations, though this theme was not explored further here. Replication of earlier results should always be an objective for researchers (Anderson & Maxwell, 2016). We extended earlier findings of high internal-consistency reliability for mixed-format exams (Garg & Lee, 2016). In the present 8 studies, we found mixed-format exam scores were consistently highly reliable (median $\omega_{ex} = .89$), across different classes, students, professors (or examiners), and examinations. Also, as expected theoretically (Dunn et al., 2014), alpha consistently underestimated reliability (estimated by ω) by about .02 for both mixed-format and PS test scores.

A limitation is that our study was confined to testing only two professors and one course in a single subject area (business finance). Further testing should answer this question. Preliminary results from other classes, courses, and professors have replicated the current results. Finally, some instructors may be unfamiliar with our recommended statistical techniques while others may view them as complex; university statisticians may be helpful.

Instructors have a wide variety of question types to use on their course exams. In this paper, we have attempted to dispel the myth that MC tests are the most reliable type. It is important since critical decisions affecting students depend on the accuracy with which marks are assigned. PS tests, at least in some courses, are highly reliable and more than MC tests. We suspect that this effect is not specific to finance, but is somewhat general, extending to other quantitative courses such as business statistics, economics, and accounting. Any professor can use our methodology to assess the reliability of student scores on their own exams.

References

Alsawalmeh, Y., & Feldt, L. (2000). A test of the equality of two related α coefficients adjusted by the Spearman-Brown formula. *Applied Psychological Measurement*, *24*(2), 163–172.

Anderson, S., & Maxwell, S. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 2016, *21*, 1-12.

Brennan, R. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*, 295-317.

Charter, R., & Feldt, L. (2002). The importance of reliability as it relates to true score confidence intervals. *Measurement and Evaluation in Counseling and Development*, *35*, 104-112.

Cox, R. (1967). Examinations and higher education: A survey of the literature. *Higher Education Quarterly*, *21*, 292-340.

DiBattista, D., and Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests, *The Canadian Journal for the Scholarship of Teaching and Learning*, 2011, *2*(2), 1-23.

Dracup, C. (1997). The reliability of marking on a psychology degree. *British Journal of Psychology*, *88*, 691-708.

Dunn, T., Baguley, T., & V. Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*, 399-412.

Ebel, R. L. (1972). Why is a longer test usually a more reliable test? *Educational and Psychological Measurement*, *32*, 249-253.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement*, *61*, 517-531.

Feldt, L. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika*, *45*, 99–105.

Frisbie, D.A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice*, *7*, 25-35.

Garg, N. & Lee, E. (2016). The reliability of mixed-format exams in higher education. In J. Doménech, J. Lloret, M. Vincent-Vela, M. Cinta., E. de la Poza, & E. Zuriaga, (Eds.), *Advances in Higher Education*. Valencia: Universitat Politècnica de València, 105-126.

Harrison, D. (2014). The uncertainty of grades in physics courses is surprisingly large. *Physics in Canada*, **70**(2), 75 (pp. 1-5).

Hill, B.J. (1978). Examination paper length: How many questions? *British Journal of Educational Psychology*, *48*, 186–195.

Hogan, T., Benjamin, A., & K. Brezinski, K. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*(4), 523-531.

Jensen, J., Berry, D., & Kummer, T. (2013). Investigating the effects of exam length on performance and cognitive fatigue. *PLoS ONE*, *8*(8), 1-9, e70270.

Lee, E.S., & Whalen, T. (2007). Synthetic designs: A new form of true experimental design for use in information system development. *ACM Sigmetrics Performance Evaluation Review*, *35*, 191-202.

Miller, M. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modelling. *Structural Equation Modeling*, *2*, 255-273.

Morley, D. (2014). Assessing the reliability of student evaluations of teaching: choosing the right coefficient. *Assessment & Evaluation in Higher Education*, *39*(2), 127-139.

Nunnally, J., & I. Bernstein (1994). *Psychometric Theory* (3rd ed.). Toronto: McGraw-Hill.

Padilla, M., & Divers, J. (2013). Coefficient omega bootstrap confidence intervals: Nonnormal distributions. *Educational and Psychological Measurement*, *73*, 956-972.

Qualls, A. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education*, *8*, 111-120.

Royal, K., & Hedgpeth, M. (2015). Balancing test length with sufficiently reliable scores. *Education in Medicine Journal*, *7* (1): e64-e66.

Williams, J., Warner, J., & Warner,S. (2004). Subject-area knowledge measured by scores on the National Association of State Boards of Geology (ASBOG) Fundamentals Examination and the implications for academic preparation. *Journal of Geosciences Education*, *5*2(4), 374-378.