1st International Conference on Higher Education Advances, HEAd'15
Universitat Politècnica de València, València, 2015
DOI: http://dx.doi.org/10.4995/HEAd15.2015.364

Mixed-format exams in higher education: Assessment of internal consistency reliability

N. Garg*, E. S. Lee*

*Sobey School of Business, Saint Mary's University, Halifax, NS, Canada B3H 3C3

Abstract: In higher education courses, instructors often use mixed-format exams composed of several types of questions such as essays, short-answer, problem-solving, and multiple-choice to evaluate student performance. It is important to discriminate reliably among students according to their performance on final examinations. The lower the reliability of student exam scores, the greater the error associated with making decisions based on them. Why then have we found no previous studies of reliability for this, one of the most common types of exam? We investigated the reliability of student scores on 12 official mixed-format final exams used in 22 classes with 1012 students in six undergraduate courses taught by five professors in three fields of business (finance, accounting, and statistics). We focussed on estimating internal consistency reliability, which is essentially a measure of the reproducibility of test scores. Using coefficient omega, the most appropriate measure for assessing reliability for mixed-format exams, we found that in these 22 classes reliability averaged .85, with over 90% of the classes with reliabilities exceeding .80. These reliabilities are very high, comparable with those reported for professionally developed standardized tests and better than those reported recently for single-format, multiple-choice exams in higher education.

Keywords: Reliability; mixed-format exams; coefficient alpha; coefficient omega; higher education; internal consistency reliability

Introduction

Professors in higher education employ a variety of different types of exams. Three of the most common types – essay only, multiple-choice questions (MCQ) only, and problem-solving only – are single-format exams with only one type of question on an exam and each question allotted the same number of marks. However, one of the most commonly used types in many academic disciplines is arguably the mixed-format exam, composed of a mixture of question types (such as short answer, multiple choice, problem solving, and essays) and with varying mark values assigned to each question (Qualls, 1995). Mixed-format exams are becoming increasingly popular even on standardized tests (Cao, 2008, p.18). These exams have the singular disadvantage of requiring an excessively long time to mark as do essay and problem solving exams (Lee et al., 2104). Nevertheless, they offer distinct advantages including ease of construction and reputedly high content validity. Given their importance in determining student success or failure, examining the reliability of student scores on examinations is of great importance.

However, "examination marks are not perfectly reliable, that is to say that if the assessment is repeated in some way, the candidate will generally receive a second mark which is different from the first" (Hill, 1978, p. 186). In higher education, it is important to discriminate reliably between students according to their final examination marks (Dracup, 1997). The lower the reliability of student exam scores, the greater the error associated with making decisions based on those scores (Crocker & Algina, 2008; Nunnally & Bernstein, 1994). Furthermore, reliability of test scores, in general, is of central importance (Henchy, 2013). Wilkinson and the APA Task Force on Statistical Inference (1999) criticized researchers for not assessing the reliability of the test scores used in their studies. Editors of many journals have argued

in a similar vein (Fan & Thompson, 2001). The same criticism can be made of classroom exams.

Reliability is generally assessed in three forms: stability over time, inter-marker reliability, and internal consistency reliability (Henson, 2001; Nunnally & Bernstein, 1994). For stability over time (Crocker & Algina, 2008, pp. 133-134), the focus is usually on assessing how student scores on an exam change over some period of time, primarily because of temporary changes in the student. Stability over time is typically estimated by test-retest reliability, the correlation between student scores on the same exam administered twice. However, test-retest reliability is of little concern here given that student exam scores on any repeated administrations of exactly the same exam would have to be suspect. Moreover, the recognition of test-retest reliability as a weak form of reliability is widespread (Morley, 2014, p. 130; Nunnally & Bernstein, 1994, p. 255). Consequently, we are not concerned with this form of reliability.

Inter-marker reliability is typically estimated by the correlation among markers in the grades awarded to students for a common exam. Various measures of inter-rater reliability have been explored by Krippendorff (2004) and Morley (2014). This form of reliability is focussed primarily on the error introduced into assessments of student exam performance by variation in how different markers score the same student exams (Crocker & Algina, 2008, p. 143). Many researchers have investigated the inter-marker reliability of classroom exams in higher education (e.g., Dracup, 1997; Hill, 1978; Newstead, 2002). In many higher education institutions, having multiple markers mark each exam in a course is economically impractical given large class sizes (e.g., in North America). More importantly, however, Morley (2014, p. 128-129) convincingly makes the case that "internal consistency is appropriate when we want to make statements about the respondent" (the student, in our case) whereas other types of reliability are appropriate for other purposes (Ebel, 1965).

Internal consistency reliability "estimates the correlation between a test and an alternative version of the same test of the same length, having randomly selected questions." Many methods have been used to estimate internal consistency reliability. However, coefficient alpha (α), which is based on the tau-equivalent measurement model (Graham 2006; Lord & Novick, 1968; Sijtsma, 2009), is the most commonly reported measure of internal-consistency reliability (Padilla et al., 2012). However, it is often an underestimation of the actual reliability because the assumptions underlying the use of coefficient alpha are frequently violated in mixed-format exams (Miller, 1995; Qualls, 1995). We argue that coefficient omega, which is based on the congeneric measurement model, provides a more accurate and more appropriate estimate of actual reliability for mixed-format exams (Dunn et al., 2014; Feldt & Charter, 2003; Schmitt, 1996). Hence, coefficient omega should be used for tests that use multiple-item formats or when the range of possible score values vary across different exam questions, as they do for the mixed-format exams in the present study (Dunn et al., 2014; Padilla & Divers, 2013a, 2013b; Qualls, 1995).

The internal consistency reliability of exams in higher education has, somewhat surprisingly, rarely been reported in the literature (Jensen et al., 2013, Cox, 1967). More recently, Jensen et al. (2013), in a quasi-experimental study of two introductory biology classes with 155 students in total, reported the internal consistency reliability of their MCQ exam to be quite poor, $\alpha = .66$. Although some professors have resorted to the use of MCQ tests in response to dramatic rises in class sizes, many eschew the use of such exams. Instead, many administer exams composed of several types of

questions (such as a mixture of short-answer, problem-solving, and essay questions) with different values assigned to each question. Yet, we could find no previous assessments of the reliability of mixed-format exams. Perhaps this is not surprising given that techniques for accurately estimating the reliability of student scores for such exams have been developed only recently (Qualls, 1995). As well, Cox (1967) noted that "although examining is an important and time-consuming occupation, very few of those who are actively engaged in it regard it as a field for experiment and research, or if they do, they keep their findings very much to themselves".

The focus of the present paper is on estimating the internal consistency reliability of mixed-format exams in a variety of classes, courses, and subject areas as well as with different professors, students, and final exams. We focused on estimating only this type of reliability, first, because it can be estimated with the single administration of a test. Second, being the most commonly reported measure of reliability (Hogan et al., 2000; Padilla et. al., 2012; Thompson, 1999), it is easily understood. Third, the other two reliability estimates, inter-marker and test-retest, were of little importance in our present studies, test-retest because it is widely recognized as a weak form of reliability (Krippendorff, 2004, p. 216) and inter-marker because in many institutions only a single individual marks each final exam. Furthermore, as Morley (2014, p. 128) pointed out "The critical difference between internal consistency reliability and inter-rater reliability is that, with the former, one is attempting to make a statement about the test-taker, and, with the latter, one is attempting to make a statement about some object of judgement such as a professor."

Method

In our study, we investigated six undergraduate courses offered at a Canadian university from three different fields in business: statistics (S), finance (F), and accounting (A). All classes were one-term 39 lecture-hour courses (see Table 1). In these six undergraduate courses, classes S1 to S15, F1, and F2 were taught in the 2nd year; F2, F3, A1, and A2 were taught in the 3rd year; and A3 was taught to graduating students in their 4th and final year of studies. The statistics and finance courses were introductory whereas in accounting the courses were either at the intermediate or advanced levels. Student performance on all exams was graded out of 100%. Each student exam was marked by a single marker (customary in many institutions) who was in all cases the course instructor. In all these courses, professors administered mixed-format exams that varied between 2.0 and 3.0 hrs in length. In these 22 classes, there were 1012 students in total. Roughly 55% were females and 45% males. One male and four female instructors, ranging from lecturers to full professors, took part in our study. A total of 12 different exams were used. For each exam, student scores on each part of each question that had been separately marked on the original exam were entered in an SPSS spreadsheet. Reliabilities were then computed for each class.

We used the MBESS program (Dunn et al., 2014; Revelle & Zinbarg, 2009; Kelley, 2007) written for the R platform for statistical computing (Field et al., 2012) to estimate reliability coefficients alpha and omega. We used the normal bootstrapping method of estimating reliabilities, as it is known to be appropriate for small sample sizes (Padilla & Divers, 2013a, 2013b).

To average our reliabilities, we used two of the methods described by Feldt and Charter (2006). In their Monte Carlo study, they examined six different approaches to

averaging internal consistency reliabilities that had been used by previous researchers. In their study, all approaches generated virtually identical averages. To be conservative, we used their approaches #1 (the simple weighted average) and #3 (the r-to-z and z-to-r transformations weighted by sample size) to average reliabilities but expected no differences between them for our data.

Results

Results of student performance in our 22 classes as estimated by reliability coefficients alpha and omega are displayed in Table 1. Both approaches to averaging reliabilities produced virtually identical values and will, therefore, not be discussed further (approaches #1 and #3 in Feldt & Charter, 2006).

Table 1. Reliability estimates for 22 classes in three higher education subjects

					Exam length		Exam marks (%)	Reliability	
Class	**Course**	**Prof**	**Exam**	**n**	**t**	**k**	**Mean (SD)**	**α**	**ω_M**
F1	Finance I	a	1	60	3.0	37	63.3 (16.9)	.85	.87
F2	Finance I	a	1	57	3.0	37	67.2 (19.2)	.86	.91
F3	Finance II	a	2	52	3.0	47	67.7 (14.9)	.85	.85
F4	Finance II	a	2	55	3.0	47	67.9 (17.2)	.88	.89
A1	Acct Int I	b	3	37	3.0	36	57.8 (14.0)	.86	.89
A2	Acct Int II	b	4	22	3.0	33	50.9 (11.4)	.67	.73
A3	Acct Topics	b	5	35	2.5	20	61.8 (15.2)	.86	.88
S1	Statistics	c	6	25	3.0	15	64.8 (23.8)	.83	.90
S2	Statistics	c	6	63	3.0	15	71.2 (22.4)	.82	.86
S3	Statistics	c	6	61	3.0	15	67.4 (23.1)	.82	.87
S4	Statistics	d	6	23	3.0	15	77.5 (14.5)	.79	.84
S5	Statistics	d	7	46	3.0	21	66.7 (19.7)	.74	.81
S6	Statistics	d	8	38	3.0	27	68.8 (15.3)	.71	.74
S7	Statistics	e	9	40	2.5	31	47.8 (15.7)	.84	.84
S8	Statistics	e	9	48	2.5	31	45.5 (19.0)	.88	.90
S9	Statistics	c	10	36	2.0	13	67.6 (19.2)	.81	.84
S10	Statistics	c	10	66	2.0	13	70.9 (20.6)	.80	.83
S11	Statistics	c	10	59	2.0	13	65.1 (21.0)	.79	.83
S12	Statistics	c	11	22	2.0	12	66.5 (26.0)	.87	.91
S13	Statistics	c	11	65	2.0	12	58.2 (23.0)	.80	.84
S14	Statistics	c	11	59	2.0	12	62.8 (22.2)	.79	.83
S15	Statistics	e	12	43	2.0	30	50.0 (19.0)	.83	.86

Note: Student marks = %; Acct Int I = Accounting Intermediate I; Acct Int II = Accounting Intermediate II; Acct Topics = Accounting Special Topics; Prof = class professor or instructor; n = number of students in class; t = maximum time allowed for exam completion (hrs); k = number of separately marked questions or parts of questions on exam; SD = standard deviation (%); α = coefficient alpha and ω_M = coefficient omega: estimated by MBESS software.

Coefficient omega averaged .85 across the 22 classes, with class reliabilities ranging between .73 and .91. The median was marginally higher at .86. Over 90% of the classes tested (20 out of 22) had reliabilities greater than .80. Average alpha equalled .82 (alphas for these classes ranged between .67 and .88). However, our empirical results confirmed the theoretical prediction that coefficient alpha underestimates actual reliability for mixed-format exams (underestimates ranged from 0 to .07). On average, coefficient alpha underestimated reliability by .035, a rather large and significant difference (sign test, 2 ties, 20/20 classes in predicted direction, $p < .0001$).

Though there appear to be some differences in the reliability of exam scores across different professors, exams, courses, and fields of study, these differences are all relatively small and inconsequential. However, we do not believe our present study

permits statistical assessments of these issues. One issue that we did address was whether student scores of shorter examinations would have significantly lower reliabilities. Professor "e" examined students using both 3.0-and 2.0-hr final exams for the same course (see Table 1). There was no significant difference between short and long exams given by this professor in this course (independent-groups $t(df = 7) = 1.47$, $p = .09$). We must caution the reader, however, of the unacceptably small sample size of only nine classes used for this test and the lack of independence of some reliability estimates (which are based on the same exam albeit for different classes).

Discussion

Professors in higher education often use exams composed of more than one type of question with variable marks assigned to each question on the exam. Many professors (mistakenly) believe that such mixed-format exams are relatively unreliable and especially poor when compared with the reliability of so-called objective MCQ exams (e.g., Cao, 2008, pp. 1 and 13). Yet we could find no reports of score reliabilities for mixed-format exams in higher education. Our study examined the reliability of student exam performance on mixed-format exams in many classes, in different courses and fields in business, and with different exams, students, and professors.

The most appropriate measure of reliability when one's focus is on decisions affecting students, as it is in our case, is unquestionably internal consistency (Morley, 2014). The most commonly reported measure of this type of reliability is coefficient alpha, but this estimate is known to underestimate the true reliability of exams composed of more than one type of question or with questions of unequal value (Dunn et al., 2014). Instead we estimated congeneric reliability using coefficient omega that is most appropriate for use with such tests as ours (Feldt & Charter, 2003; Qualls, 1995).

Reliability of exam scores was very high with coefficient omega averaging .85 in the 22 classes in our study. Moreover, the reliabilities were remarkably consistent from class to class (ranging from .73 to .91) despite variation in students, professors, exams, fields of study courses, and classes taught. Scores on the exams in over 90% of the 22 classes tested in our study had reliabilities exceeding .80. Practically speaking, these reliabilities imply that error is relatively small, and decisions based on student performance on this type of exam in the courses tested are well founded.

However, decisions on students are often based not just on their performance on a single final exam but on assignments, midterms, projects, and presentations in the same course as well. Such additional measures of student performance generally increase reliability (Nunnally & Bernstein, 1994). Thus, reliability of student scores on a single final exam undoubtedly underestimates the reliability of student grades assigned for all aspects of a course. Furthermore, decisions about students are often made on the basis of performance in many different courses with different professors, fields, class sizes, and time periods. As Dracup (1997) has shown, reliability based on student performance on essay-only exams in many courses can be exceptionally high ($\alpha = .95$) even when the (inter-marker) reliability of most courses is very poor (median inter-marker reliability = .64 with some reliabilities as low as -.28).

Previous studies of reliability are relatively rare in higher education (Cox, 1967; Jensen et al., 2013). Recently, however, Jensen et al. (2013) reported $\alpha = .66$ for the internal consistency reliability for student scores on a single MCQ exam in two introductory biology classes taught by one instructor. Such poor reliability implies that error is

relatively high and that decisions based on the results of such MCQ exams could be somewhat compromised. The reliability for all 22 of the classes in our study exceeded that for their MCQ exam scores. The reliabilities we have found for mixed-format scores compare favourably with those found by others for MCQ and other types of exam scores in higher education. In fact, the reliabilities in Table 1 are comparable with those reported for costly, professionally-developed standardized clinical and psychological tests. Yet, classroom exams, such as ours, are normally intended for one-time use (Nunnally & Bernstein, 1994, p. 295).

Several limitations should be stressed. First, in this paper we focussed exclusively on investigating internal consistency reliability to the exclusion of other forms such as inter-marker reliability, which explore different sources of error (Crocker & Algina, 2008). However, as Morley (2014, p. 128) so clearly affirms, internal consistency reliability is useful for making judgments about students while inter-marker reliability is useful for making judgments about professors. Given our focus on the importance of marks or grades on exams when making decisions about students, internal consistency reliability is unquestionably most appropriate. Similarly, the issues of moderation, a method for improving marker consistency in which several markers meet to iron out differences, and calibration, in which markers learn to mark more consistently by working with other markers, are both concerned with inter-marker reliability (Sadler, 2013), and therefore, while important issues in their own right, are not the focus here.

Second, our interpretation of the present results must be tempered somewhat by the relatively small number of exams, instructors, classes, courses, and subject fields tested in the present study. Nevertheless, we surveyed 1012 students, 12 exams, 5 instructors, 22 classes, 6 courses, and 3 subject fields. In forthcoming studies, our objective is to cover more subject areas, courses, classes and students, exams, and professors.

Third, we studied only mixed-format exams. Nevertheless, these are among the most commonly used types of exams in higher education. Others have reported reliabilities for essay-only exams in psychology (e.g., Dracup, 1997), problem-solving-only exams in engineering (e.g., Hill, 1978), and MCQ exams in biology (e.g., Jensen et al., 2013). Those interested in these other types of exams should consult the references cited.

Fourth, are the exams in these 3 disciplines really comparable? This is an important issue which cannot be addressed in appropriate depth here given space constraints. Nevertheless, several arguments can be advanced for believing that at least some of the exams are comparable. All courses examined in our study are, for example, in applied disciplines (e.g., accounting). All exams included both strong quantitative and strong non-quantitative or narrative components. Each mixed-format exam also included many problem-solving and narrative short-answer questions. Nevertheless, another response to this question is that they are certainly not. Questions on finance exams bear little resemblance to those on statistics exams. Even within a discipline, exams on introductory courses can differ radically from those on more advanced courses. However, readers will certainly differ in what they believe constitutes exam comparability. This is why the reliabilities are provided for each exam in each of the classes, courses, and disciplines studied (see Table 1). Finally, one must consider that, despite these manifold differences, reliabilities for these widely divergent exams were uniformly high in our study (more that 90% had reliabilities exceeding .80).

In our forthcoming studies, our objective is to cover more subject areas, courses, classes and students, exams, and professors. The present studies were restricted to an exploration of internal consistency reliability only. Other estimates of reliability such

as inter-marker reliability, which explore other sources of error, were not addressed (Crocker & Algina, 2008). While we have explored reliability for classes in three areas of business, other areas such as economics were not examined in our study. Similarly, we did not investigate courses in arts, sciences, or engineering. Additionally, we have not studied single-format exams such as those consisting of only essay questions.

References

Cao, Y. (2008). *Mixed-format test equating: Effects of test dimensionality and common-item sets.* (Doctoral dissertation, Department of Measurement, Statistics and Evaluation, University of Maryland, 2008).

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

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Fort Worth: Harcourt, Brace and Jovanovich.

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

Dunn, T., Baguley, T. & 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. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice-Hall.

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. S., & Charter, R. A. (2003). Estimating the reliability of a test split into two parts of equal or unequal length. *Psychological Methods, 8(1),* 102.

Feldt, L. S., & Charter, R. A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement, 66(2),* 215-227.

Field, A., Miles, J., & Field, J. (2012). *Discovering statistics using R*. London: Sage Publications, Ltd.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability what they are and how to use them. *Educational and Psychological Measurement, 66(6),* 930-944.

Henchy, A. M. (2013). *Review and evaluation of reliability generalization research*. (Doctoral dissertation, Educational, School, and Counseling Psychology, University of Kentucky, 2013).

Henson, R. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34,* 177- 189.

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

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (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.

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods, 39(4),* 979-984.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Communications Research, 30(3),*411-433.

Lee, E., Bygrave, C., Mahar, J., Garg, N. and Cottreau, S. (2014) "Can exams be shortened? Using a new empirical approach to test in finance courses", International Conference on Higher Education and Management (ICHEM 2014), London, UK, January 2014.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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.

Newstead, S. (2002). Examining the examiners: Why are we so bad at assessing students? Psychology Learning and Teaching, 2(2), 70-75.

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

Padilla, M. A., Divers, J., & Newton, M. (2012). Coefficient alpha bootstrap confidence interval under nonnormality. *Applied Psychological Measurement, 36(5),* 331-348.

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

Padilla, M. A., & Divers, J. (2013b). Bootstrap interval estimation of reliability via coefficient omega. *Journal of Modern Applied Statistical Methods, 12,* 78-89.

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

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74(1),* 145-154.

Sadler, R. (2013). Assuring academic achievement standards: from moderation to calibration. Assessment in Education: Principles, Policy and Practice, 20(1), 5-19.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350-353.

Sijtsma, K. (2009) On the use, misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107-120.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory and Psychology, 9,* 165-181.

Wilkinson, L., & the APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist, 54,* 594-604.