Can Higher Education Exams Be Shortened? A Proposed Methodology

Eric S. Lee, Connie Bygrave, Jordan Mahar, and Naina Garg

***Abstract*—Any lecturer would agree that marking exams is the bane of her existence. A time-consuming and tiring process, it often requires complex, subjective judgments. Higher education exams typically take 3.0 hours. Do they really need to last so long? Can we justifiably reduce the number of questions on them? Shortening an exam by one hour, if justified, should result in a one-third reduction in lecturer time and effort spent marking. Surprisingly little empirical research has addressed these problems. Classical methods may be partly to blame for this dearth of studies. We propose an alternative methodology based on three key components including two recent developments in experimental design and statistics -- synthetic experimental designs and equivalence hypothesis testing. The third component consists in comparing, on six psychometric criteria, student performance in a class on the standard 3.0-hr final exam with that on shortened exams with proportionately fewer questions. Two are the frequently misunderstood standard psychometric criteria – reliability and validity. We argue that adding four common-sense criteria – justifiability of test use, number of exam questions, equivalence in mean student performance, and correspondence (between shortened and full-length exam scores) – confer significant additional benefits. Our approach provides a simple methodology that lecturers can, with minimal time and effort, use to examine the effect of shortening exams for their own classes.**

***Index Terms*—Exam length, psychometric criteria, synthetic experimental designs, test length.**

I. Introduction

Marking exams is the bane of any lecturer's existence. It is characteristically a tedious, time-consuming process. The intricate subjective judgments that are required can exhaust even the most dedicated of lecturers. Yet despite the almost universal loathing of lecturers for this activity, surprisingly little research has been published on ways to reduce the time and effort required to mark conventional written examinations. The purpose of the present study is to redress this long-standing neglect. In this paper, we are concerned with mixed-format exams used most commonly in academe and consisting of a mixture of different types of questions including problem solving (requiring detailed solutions), essay, short answer, and multiple choice questions.

Written final examinations three hours or more in length are common in many universities and colleges around the world. Why then, if marking is so universally loathed, are exams so long (both in duration and number of questions posed)? Custom or tradition seems the primary reason [1]. Even more surprisingly, long exams have been retained despite dramatic increases in class sizes over the past couple of decades at most institutions of higher learning. Many lecturers complain of class sizes more than doubling in just the last 15-20 years [2]-[4]. In effect, marking time and effort for many lecturers has, at a minimum, effectively doubled. As most lecturers already work near their limits (far more than the cultural norm today of 35-40 hours per week), this dramatic increase in marking necessitates a comparable costly reduction in other activities such as research or mentoring students. Official policy at many institutions also increases the pressure on markers. Final grades must often be submitted within an unrealistically brief period.

In response to such pressures [4], some lecturers have abandoned written examinations in favour of multiple choice question (MCQ) exams. The latter have been extensively studied and will not be reviewed here. We note in passing, however, that although MCQ exams offer many advantages (e.g., no effort required to mark them), they are, in the opinion of many lecturers, unsuitable for use in many university and college courses. For example, smaller class sizes, rapidly changing course coverage, or advanced topics are often better handled using mixed-format written examinations [5], [6]. Furthermore, most lecturers are unaware of and untrained in the extensive theory, proper design, and use of MCQ exams. It takes a lot of time and very hard work to design effective MCQ exams, time that, in our experience, is rarely devoted. Sadly, even the banks of multiple choice questions provided with many introductory textbooks have often been poorly prepared. Given the many shortcomings associated with the use of MCQ exams, many of us prefer to continue using traditional mixed-format written examinations.

Common sense, however, suggests the time required to mark mixed-format written examinations can be substantially reduced by shortening the length (that is, by reducing the time allotted for exam completion and proportionately reducing the number of questions posed). An obvious caveat is that the shortening should not destroy the psychometric properties of the exam, that is, the suitability of the exam for assessing student learning and knowledge in an academic course.

A. Empirical Studies of Exam Length

Over 50 years ago, Cox [7] observed wryly "It is, however,

Manuscript received November 6, 2013; revised January 12, 2014. This work was supported in part by Fairleigh Dickinson University and by Saint Mary's University.

E. S. Lee and N. Garg are with Finance, Information Systems, and Management Science Department, Saint Mary's University, Halifax, NS, B3H 3C3 Canada (e-mail: elee@smu.ca).

C. E. Bygrave is with Masters in Administrative Sciences Department, Fairleigh Dickinson University, Vancouver, BC V6B 2P6 Canada (e-mail: bygrave@lfdu.edu).

J. Mahar is with Mathematics Department, Dalhousie University, Halifax, NS, B3H 4R2 Canada.

of special significance here, since, 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." Not much has changed in the 50 years since he wrote those words.

Nevertheless, there are a few. However, we leave comprehensively reviewing the past literature on this topic for a later paper. Here we confine our discussion to noting that almost all of these articles have restricted their attention to assessing the reliability of exam scores [1].

The typical approaches that have been used to research issues of test length are discussed in classic psychometric texts such as [8]. Typically the focus is on estimating the reliability of test scores and, more rarely, test validity [9]. Two recent papers typify the approaches frequently used to investigate these issues. To study the effect of different SAT test lengths on performance and subjective fatigue effects, [10] employed a conventional independent-groups experimental design while [11] employed a conventional dependent-groups experimental design. While both of these study designs were appropriate to examine the specific questions addressed by these authors, they are also time-consuming, complex, and difficult to conduct. These traditional experimental approaches are also problematic if used to examine issues of classroom exam length.

Another researcher, Hill, pioneered an alternative approach [1]. Our approach, though based on Hill's method, differs significantly from his, as we shall discuss at length in the following sections. We believe that our recommended changes greatly increase the advantages of Hill's method resulting in a much more effective, simpler, and less time-consuming methodology for studying the problem of higher education exam length.

II. The Classical Psychometric Approach

For over 100 years, classical test theory has been predicated on the proposition that test development must be based on two key criteria -- reliability and validity [12].

A. Reliability

To be useful, the performance scores of individual students on some class examination must be reliable. As Dracup [9] asserts, "Reliability is a fundamental requirement of any assessment procedure. The greater the reliability of an assessment, the more certain we can be that observed differences between the individuals on the assessment are the result of real differences between the individuals on whatever the assessment is measuring rather than the result of random error." Reliability and measurement error are inversely related as the error associated with student scores on an exam generally decreases as reliability increases. Error or unreliability of student marks can be caused by a multiplicity of factors, including the particular questions sampled and marker error. Frisbie [13] expressed it well: "We should not expect test scores to be perfect measurements, but there is only so much error that we should be willing to tolerate. A geography test should yield scores that put students in the same relative order, whether given on Tuesday or Wednesday. If a slightly different set of 40 test questions had been used, essentially the same relative ordering of student scores should have resulted. That is, anytime a classroom test is given, we would like the resulting scores to be generalizable over testing occasions, over sets of similar test questions, and over slightly varying test conditions. … If we cannot rely on the scores as accurate measurements of achievement, we cannot use them to make instructional decisions or to communicate progress to students".

It should be apparent that "reliability is a property of a set of test scores, not a property of the test itself" [13].

B. Validity

Arguably the most important criterion that must be satisfied when considering alternative exams is validity, or the degree to which they measure what they purport to measure. To be more specific, "validity is defined not strictly as a characteristic of a test or as a characteristic of a score but necessarily as an interaction of the two" [14]. Furthermore, "[v]alidity is the degree to which scores on an appropriately administered instrument support inferences about variation in the characteristic that the instrument was developed to measure" Validation, on the other hand, is defined as "the ongoing process of gathering, summarizing, and evaluating relevant evidence concerning the degree to which that evidence supports the intended meaning of scores yielded by an instrument and inferences about standing on the characteristic it was designed to measure". Claims and decisions based on official shortened exams should, therefore, be as valid as those based on current full-length 3.0-hr exams.

C. Critique

Despite the undisputed importance of reliability and validity as criteria that must be satisfied in test or exam development, we argue that this classic approach suffers several weaknesses. First, reliability and validity are difficult, complicated concepts understood by few of their users and misunderstood by many [13]-[15]. Second, very often only reliability is reported in the literature. Does this reflect a general reluctance among researchers to deal with the subjectivity inherent in any investigation of validity? Third, we argue that there are other issues, ignored by classicists, that are relevant and of some importance in developing effective tests and exams. These issues–justification of test use, number of questions on exams, correspondence, equivalence – are discussed in subsequent sections. Fourth, these other issues are, we argue, based on common sense and more easily understood. Furthermore, all six criteria can be used profitably in comparing psychometrically the effectiveness of shortened exams with full-length exams.

III. Our Proposed Approach

A. Purpose

We propose a methodology that professors can use to ascertain whether there is empirical support for justifying a marked reduction in the length of their current final examinations, or whether, as Hill found, the evidence would indicate the need to increase current exam length [1]. For

each class in a course of interest, student performance on one or more shortened exams can be compared with that on the full-length 3.0-hr exam. Shortened exams in each class would have proportionately fewer exam questions. Even a one-hour reduction in exam length from the traditional 3.0 hours to 2.0 hours, if justified, should result in a substantial one-third reduction in the total amount of time and effort required to make up the exam and mark it. To compare shortened with full-length exams, a separate experiment would be conducted on each class. The same experimental procedures are employed with each class.

B. Exams

With our proposed method, each student actually writes only a single final exam, the current full-length exam. For illustration purposes, we will assume in the ensuing discussion that the current exam is 3.0 hours long, the typical length of a final exam at our university and the shortened exam is 2.0 hours in length.

C. Experimental Design

We considered conventional independent-groups and repeated-measures experimental designs for examining the effect of exam length on student performance but rejected both methodologies. For example, a conventional independent-groups experimental design would have required students in a given class be randomly assigned to write one of two possible final exams (a 2.0 and a 3.0-hr exam). Similarly, a conventional repeated-measures design would have required all students in a given class to write both exams with the order in which each student wrote the two exams determined randomly. Clearly, both types of traditional randomized experimental designs must be rejected for use here given the many obvious deficiencies associated with each: practical difficulties, ethical concerns, cost, and time [16]. For example, both students and administrators would refuse to accept the use of multiple final examinations of varying lengths for any course. Universities and colleges generally require that for the sake of fairness all students in a given course be evaluated in the same way. As well, picture the poor student compelled to write two exams totaling 5.0 hours during a busy exam schedule in which they had to write final exams in 3-5 other courses! Similarly, pity the administrator confronted by irate students complaining of unfair treatment because they had been randomly assigned to write a long 3.0-hr final exam while other students in the same course only had to write an easier 2.0-hr exam!

Instead, to get around these manifold problems, we propose employing the same experimental methodology pioneered by Hill [1] for investigating the effects of exam length on student performance and more fully explored theoretically in [16]. The latter authors refer to this variation on traditional repeated-measures designs as a synthetic experimental design. Synthetic designs are true experimental designs capable of examining cause-effect relationships and with many of the advantages of both conventional independent- and dependent-group designs and few of their respective inherent disadvantages. As will be seen in subsequent sections, synthetic designs offer marked practical and ethical advantages as well as simplicity and low cost.

Synthetic designs are a form of repeated measures (or dependent-groups) in which the effects of a variable are examined experimentally using synthetically generated performance scores for each subject for all comparison groups except the original set of empirical performance scores. For example, one can use student performances on a full-length 3.0-hr exam, the empirical scores, to generate the synthetic performances on shortened comparison exams.

The methodology employed is virtually identical in each class (i.e., experiment) and is described in the following sections. For each class, a synthetic experiment would be conducted with one synthetic factor, or independent variable - exam length.

Synthetic designs, unlike conventional experimental designs, do not require that an empirical study be designed and then conducted to gather the information required to address the research issues of interest (though this can also be done). In fact, archival data can be used as the basis for a synthetic experimental design [16]-[21]. Hill [1] was the first to use a synthetic design to investigate the issue of exam length though he did not take advantage of the many uses that can be made from such an experimental design [16]. This is, in fact, what was done here. We argue that many educators and researchers could benefit from using this variation of a true experimental design.

D. Procedure

Students actually write only a single exam. For present purposes, we assume that to be 3.0 hours long. Synthetic experiments are always conducted in two successive phases: an empirical phase followed by a synthetic phase [16].

1) Empirical phase

No new data need actually be collected empirically for the purposes of this study. Instead, all student exams would be removed from storage (regulations at most universities dictate storage of all final exams for a year in case there are appeals), and for each student, the marks awarded for each part of each question that had been graded separately on the full-length 3.0-hr exam would be recorded in a spreadsheet, e.g., Excel or SPSS. Thus, an empirical data set would consist of the set of n student vectors in a class, each vector composed of the marks awarded to that student on each part of each question (that could be separately scored) on the full-length exam.

2) Synthetic phase

For the purposes of this illustrative example, the original exam writer (i.e., examiner) of a given full-length 3.0-hr exam would subsequently construct shortened exam versions of the desired lengths. The same procedure is used to construct each shortened-length exam. The subset of questions on any shortened exam always constitutes a subset of all the questions on the full-length exam. The subset should always be selected by the examiner to produce the best possible, fairest exam (and appropriate for the time available) of the shortened length given this constraint. As a rough guide in each case, the ratio of exam times (short/full) sets the percentage of marks for questions selected from the full-length exam to be included on the newly created shortened exam. Thus, for a 1.5-hour exam, for example, a subset of questions would be selected totaling approximately

(1.5/3.0 × 100 =) 50% of the original 100 marks allotted on the full-length exam.

A spreadsheet equation can be used to compute the mark a student would receive for each version of the 3.0-hour exam without resorting to further empirical testing of students. Each equation sums the marks achieved by a student on only those questions (or parts of questions) that would appear on that particular shortened version of the full 3.0-hour exam. In a few cases, the mark allotted for a particular part of a question can be changed slightly (by at most a few per cent) to reflect the relative importance of the question on the new shortened exam or to make the questions sum to the desired target.

For comparability of exam performances, student marks for each shortened exam are then renormalized to a range of 0 to 100%. Thus, if the total marks on a shortened 1.5-hour exam add up to 48% (of the marks on the original 3.0-hr exam), then each student's shortened exam mark is multiplied by 100/48. A student mark of 36 out of a maximum possible 48 on such a 1.5-hour exam would, therefore, result in a score of 75%.

Generation of derived, or synthetic, student performance scores are based on the assumption that students would answer the same question in exactly the same way on a shortened exam as they had on the actual full-length 3.0-hour exam. Given that a comparable amount of time would be available to answer this identical question in both exam situations, this assumption seems reasonable. In his early investigation of the problem of exam length, Hill [1] generated synthetic performance scores for students on various hypothetical shortened engineering exams. Though he did not explicitly state it in his paper, this same assumption must be made to justify the statistical analyses and conclusions that he made. Similarly, in their recent searches for effective shortened versions of the Raven Advanced Progressive Matrices test, [17], [18], [21], also synthetically generated performance scores for shortened versions of this test. Though not stated explicitly, a similar assumption must be made to justify the generation of synthetic empirical data, the statistical analyses, and the interpretations and conclusions that these authors made. This assumption must always be made to justify use of a synthetic design [16].

E. Criteria for Evaluating Suitability of Shortened Exams

Our approach to assessing whether a shortened version of a traditional 3.0-hr final exam could replace it as the official final exam in a course is based on six psychometric criteria: 1) reliability, 2) validity, 3) justifiability of test use, 4) the number of (separately scorable) questions on an exam, 5) correspondence (between the performance of students on the full-length 3.0-hr exam and that on a shortened exam), and 6) equivalence of, and differences between, mean student performance on shortened and full-length exams. Reliability and validity are traditional psychometric requirements that should be met by any examination that is used to assess student performance in a course [8]. They have usually been the only psychometric properties examined when developing standardized achievement, diagnostic, and counseling tests.

We propose consideration of four additional psychometric properties–justification of test use, number of exam questions, equivalence, and correspondence–because they provide easily understood, comprehensive insight into the assessment of the suitability of shortened exams as replacements for current 3.0-hr exams. These four criteria are new ones that should also be met to justify shortening any current exam. We do not claim that our six criteria provide completely independent sources of psychometric information relevant for decision making, just as classical notions of reliability and validity do not. Rather we argue that they provide a more comprehensive, clearer picture of the strengths and weaknesses of any proposed shortened test. We discuss each criterion in turn, explaining why we use each, how we measure each, and the standard we set to be met by any shortened exam to be considered a suitable replacement.

1) Reliability criterion

Though there are many types of reliability (such as test-retest, intermarker, intramarker), we focus on internal-consistency reliability for three reasons. First, it can be assessed from a single administration of a test. Second, it is the most frequently reported measure of reliability [8], [22]. Third, we are most interested in estimating, not intermarker reliability or test-retest reliability, but the reliability of interpretations of student scores on our exams (i.e., measurement error primarily due, not to differences among different markers, but to variation among the items on an exam). "Internal consistency estimates relate to item homogeneity, the degree to which the items on a test jointly measure the same construct" [23]. The questions or items on an exam should be intercorrelated, thereby permitting lecturers to interpret meaningfully the sum of marks for all questions on an exam.

Following common psychometric practice, we estimate internal consistency reliability using coefficient alpha (α) which estimates the correlation that one would expect between a test and some alternative version of the same test of the same length, having the same number of randomly selected questions [8], [23]. This can also be thought of as the correlation one would expect between an actual test and errorless true scores. For example, if α = .81, then the expected correlation between an actual test and errorless true scores would equal .90. Alpha is like a correlation coefficient, with larger values signifying higher reliability and ranging from 0 to 1.00. According to Frisbie [13], "Alpha can provide an estimate of the reliability of scores from tests composed of any assortment of item types – essays, multiple-choice, numerical problems, true-false, or completion."

It is important to remember that α underestimates the true reliability [24]. To be precise, "the theoretical reliability coefficient can be characterized as the coefficient of precision (i.e., the correlation that would be obtained between two perfectly parallel forms of the test if there were no changes in examinees between testings). When a composite is made up of nonparallel subtests, we can estimate the lower bound of its coefficient of precision by using coefficient alpha," [25].

What reliability (estimated by α) should we expect of any acceptable exam, whether shortened or full-length, used to assess student performance in a course? Two standards have conventionally been applied when assessing reliability using

coefficient alpha. Many researchers cite Nunnally & Bernstein [8] as advocating the use of $\alpha \geq .70$. However, a careful reading of [8] does not support this position [21]. Moreover, the $\alpha \geq .70$ standard was intended for use by researchers, not those making decisions on the basis of class marks on exams. Others have advocated the use of $\alpha \geq .90$, but this standard is most appropriate for use in the development of standardized achievement and diagnostic tests because such a test is often used by itself to assist in making important decisions affecting the future of people. We suggest that readers appreciate that the standard for the level of reliability to be achieved should depend upon the uses to which scores on the test are to be put [13], [8]. Most decisions in higher education are based on student performance in multiple academic courses. The reliability of combinations of course marks is very, very high, often exceeding .90 [9], even when the reliability of marks in each individual course is very low.

In general, there is no magical cut-off that should be exceeded, and we argue that there is certainly none that can be defended for use in addressing issues in exam length [13]. As Schmitt [26] states so succinctly, "There is no sacred level of acceptable or unacceptable level of alpha. In some cases, measures with (by conventional standards) low levels of alpha may still be quite useful." Similarly, Schmitt asserts "When a measure has other desirable properties, such as meaningful content coverage of some domain …, this low reliability may not be a major impediment to its use" (as would hopefully be true of university exams). Nevertheless, the higher the reliability, the better it is for making decisions about students.

However, Fan & Thompson [27] argue that in any study a standard for reliability should always be specified and a rationale provided for that value. Based on our long experience teaching, our knowledge of psychometric principles, and the quantitative nature of the courses that we teach at this time (statistics and finance), we argue that alpha should equal or exceed roughly .75 ($\alpha \geq .75$), with the caveat that even higher reliabilities are preferable. Lower standards might be more appropriate in other types of courses such as non-quantitative subjects (e.g., English). Our rationale for this criterion is based on three considerations. First, reliability for quantitative courses such as statistics and finance are undoubtedly generally higher than that for non-quantitative courses [9]. By nature, quantitative disciplines such as statistics and finance are more objective than non-quantitative ones such as organizational behaviour. Finance exams test student ability to apply formulae and theories to exact the "right" answer. Organizational behavior exams test student ability to apply theories to critically analyze situations. The "right" answers are contingent upon the examiner's judgment. This subjectivity can affect the reliability of test scores.

Second, we note that reliability for a final exam in a course need not be so high (as .90) since final grades for students in North American education institutions are typically based not just on student performance on the final exam, but also on performance on midterms, assignments, projects, and others [13]. These additional sources of evidence on student performance in a given course can increase the reliability of the overall final grade substantially.

Third, decisions affecting students are typically based on their performance in many different courses. If we accept that ideally the overall reliability for a collection of many courses (on which important decisions affecting a student are to be made) should be very high (e.g., $\alpha \geq .90$), then the reliability for any particular course could be much lower than .90. As well, if the overall student mark in a course is based on other factors such as midterms in addition to final exam performance, then the ideal reliability standard of the final exam scores by itself could be set even lower still. To be on the conservative side, however, we set the minimum standard to be exceeded for reliability of exam scores in a given class of our quantitative courses as $\alpha \geq .75$ or .80.

2) Validity criterion

We believe that three sources of validity evidence for the interpretation of student performance on exams can be examined relatively easily: face validity evidence, evidence based on content validity, and evidence based on internal structure [8], [15]. This is one more than the modal number of sources commonly reported in research articles aimed at establishing the sound psychometric properties of achievement, psychological, and counseling tests [15]. We define each and describe how each might be measured.

Face validity refers to the degree to which examiners and students subjectively judge an exam to be fair, reasonable, and appropriate, that is, how well does the exam cover the knowledge and skills taught in the course. Given our focus on the professor teaching a course, one could ask whether or not they would be willing to use each shortened exam version as the official final exam in their course. As well, we could ask them to rate on a 5-point Likert scale (1 = not at all acceptable to 5 = very acceptable) the acceptability of each exam as the official exam for the course. Though alternative methods of estimating face validity can be used (e.g., asking a panel of experts), they are generally more time-consuming.

Content validity refers to how well the questions on a given exam sample the content covered in a course. The typical approach for establishing evidence for this source of validity is by use of an effective method for ensuring content validity in the construction of the exam [8]. For exams, therefore, constructing a comprehensive plan, outlining the topics taught, their relative importance in the course, the amount of class time spent on each topic, and the type of student being examined is often a more effective method of achieving validity than conducting an investigation of the content validity achieved with an already constructed exam. This approach should be used for construction of all official exams but would be inadmissible for shortened versions. For shortened exams, therefore, one could rely on assessing content validity using a short questionnaire administered after creation of the shortened versions of each 3.0-hr exam.

To measure content validity, one might consider asking the professor teaching a course two questions. First, how well did each version of an exam cover all important topics in the course? Second, how well did the mark allocation on each exam reflect the relative importance of the topics covered in the course (i.e., were more marks allocated to more important topics in the course). Each question could use the same 5-point Likert scale of 1 = not very well to 5 = very well.

Evidence based on test content is most appropriate for assessing the validity of higher education course exams.

Internal structure validity refers to the relationship among the questions on a test or exam. Two possible measures of internal structure evidence are coefficient alpha for each exam and the item-to-total correlation for each question on each exam. Contrary to popular belief, internal consistency reliability does not assure validity [28]. It does, however, set an upper bound on the possible validity associated with an exam. Consequently, high internal consistency reliability is crucial if one hopes to develop effective, content valid exams. Therefore, we suggest examination of reliabilities (as estimated by α) associated with various exams to assess evidence for validity based on internal structure.

3) Justification-of-use criterion

Many have argued about the general lack of understanding of reliability and validity [13], [14]. To address this, we follow Cizek's [30] lead in insisting that "A distinction must be made between evidence supporting the intended inferences from test scores and evidence supporting a test use. Validity theory must be refined to differentiate between validation of score inferences (i.e., the methods and sources of information relevant to determining the confidence that is warranted regarding the intended meaning of a test score) and justification of test use (i.e., the methods and sources of information – including consequences – brought to bear on the question of whether it is a good idea to use a test in the first place)."

There are, however, no generally accepted sources or standards of justification evidence. Nevertheless, we believe that four sources of justification can be considered when assessing this criterion for shortened exams [29]. First, what are the consequences of using a shortened exam in place of the full-length 3.0-hr exam as the official test for a course? Second, what changes in the human and financial resources and costs can be expected by adopting a shortened exam? Third, could other policy goals be achievable by a proposed shortening (e.g., more research or more attention to mentoring students)? Fourth, what are the relative benefits of using a shortened exam as the official final exam for a course instead of the current 3.0-hr exam length?

4) Number of exam questions criterion

The number of (separately scorable or gradable) questions, or parts of questions, on an exam is related to both reliability and validity. The Spearman-Brown prophecy formula (which can be derived from classical test theory) indicates how, provided the assumptions of classical theory are met, reliability (in our case, as estimated by coefficient alpha) can be increased, or decreased, by simply increasing, or decreasing, the number of questions on an exam. This equation clearly shows that any exam writer has only to increase the number of questions on an exam (but keeping exam length in time constant) to secure higher reliability (and to increase simultaneously the upper bound for validity) [8].

A second advantage of explicitly considering the number of questions on an exam is Nunnally and Bernstein's [8] admonition to ensure that a minimum of at least 10 questions should appear on any test or exam to secure adequately high reliability (i.e., ensure $k \geq 10$). The investigation by Hill [1] shows the value of this rule. If Hill had followed this rule, he would have realized that none of his full-length 3.0-hr exams had a sufficient number of questions. With only 5-6 questions on each of his 3.0-hr engineering exams, it is clear that reliability could not be high enough to justify shortening the then-current official 3.0-hr exams. In fact, more, rather than fewer, questions should have been asked on each of his official 3.0-hr exams.

A third potential advantage is that it immediately suggests how, if reliability is low enough to be of concern, one need only consider increasing the number of questions posed on an exam (for a given length of exam time).

5) Correspondence criterion

Another common-sense criterion that should, in our opinion, be met by any suitable shortened (replacement) exam is that students in a class should perform as poorly (or as well) on a suitable shortened exam as they did on the official full-length 3.0-hr exam. That is, we expect student marks on a suitable shortened exam to correspond highly with that on the full-length 3.0-hr exam. Pearson r can be used as a measure of correspondence. While reliability also addresses this issue, this criterion provides an alternative way of looking at this issue. It is analogous to the criterion employed by researchers developing shortened forms of already established full-length psychological tests [17], [18], [21], [30].

As discussed earlier, practical and ethical difficulties make it almost impossible to assess this criterion using conventional experimental methods. Synthetic experimental designs, however, make this possible. For each student in a class, synthetic designs make it possible to generate scores for students on shortened versions of the original exam. Because synthetic designs are a form of repeated measures, they also provide a very powerful test of this criterion.

As a standard, we expect that the correlation between the full-length exam and an acceptable shortened exam would exceed roughly .90 (i.e., $r \geq .90$) so that at least 80% of the total variation among student scores on the full-length 3.0-hr exam scores can be explained by variation in shortened exam scores. We considered shortened exams with lower r values to contain too much error to be used as replacements for current full-length exams.

6) Equivalence (and difference) criterion

A final common-sense criterion is that student performance on a suitable shortened exam should, on average, be roughly equivalent to that on the current full-length 3.0-hr exam. Most professors who have taught the same course several times recognize that average student performance varies from one class to another. In part, this is expected since the exams are never the same from one class to another, the students are not the same, and how one teaches changes from one year to the next. Nevertheless, most professors would agree that these class averages, while never exactly the same, should, if the exams are fair, be roughly equivalent. For a shortened exam to be considered an acceptable substitute for the full-length exam, we hypothesized, first, that average student performance on that exam would not deviate significantly from that on the full-length 3.0-hr exam, and second, that average student performance on the two exams

should be roughly equivalent.

However, traditional hypothesis tests, sometimes referred to as difference tests, cannot test for equivalence. Many researchers mistakenly believe that a failure to reject the null hypothesis (H_0) in difference testing provides evidence that no real difference exists between groups. Such a conclusion has been shown repeatedly to be wrong [31]. In reality, the only justifiable conclusion one can make on failing to reject the null hypothesis is that there is insufficient evidence of a real difference existing.

In recent years, statisticians in the medical and pharmaceutical fields have developed an alternative to the conventional difference test method of testing hypotheses. In equivalence testing, the experimental or alternative hypothesis (H_1) directly tests for equality of means. In this method, the null hypothesis (H_0) is that there is no evidence of equivalence among the group means. Readers should note that in equivalence testing a conclusion of H_0 does not signify that a difference exists. To test whether a difference exists among groups, one must use difference testing. Similarly, to test whether groups are equivalent, one must use equivalence testing. We follow the lead of [32], [33] in using both difference and equivalence statistical tests.

F. Statistical Analysis of Exam Length Data

In a separate paper, we discuss the important issues associated with the statistical analysis of exam length data collected using our approach [34]. We restrict our comments here to note that in that paper we present recommended procedures to deal with important issues in statistical analysis such as equivalence testing, problems with conventional hypothesis testing, and confidence intervals for correlations, reliabilities, and differences between means.

G. Estimating Equivalence in Mean Grades on Exams

Equivalence testing requires the estimation of delta (Δ) which determines the range ($\pm\Delta$) within which the observed difference between mean student performances for a class between shortened and full-length exams could normally be expected to fall. We believe delta can be estimated using any one of four different methods, two providing subjective estimates and two empirical. The first method consists in asking the instructor in a course to estimate the typical range within which mean student marks of previous classes had varied. The second method consists of asking a sample of lecturers who have taught the same course before to estimate subjectively the range within which mean student marks for different classes and different full-length 3.0-hr exams would normally be expected to fall (and still be considered to be fair measures of student performance in a course).

The third method consists in empirically estimating delta by examining the mean final exam mark for a sample of previous classes taught that all had full-length 3.0-hr exams. A rationale for using this approach is that previous classes with 3.0-hr exams judged to be roughly equivalent measures of student performance in the course would provide a reasonable empirical estimate of delta. Since a lecturer has judged all class means in their course to be essentially equivalent, despite the manifold differences among the exams used in each class, in the students taught, and when they were taught, delta can be estimated to be approximately equal to roughly two standard deviations away from the mean of the class means observed in the recent past in a given course for 3.0-hr exams (for a 95% confidence interval).

The fourth method of estimating delta would be to set $\Delta = \pm 10\%$ of the control group mean for each class (that is, of the 3.0-hr exam class mean). Thus, if the control group is the 3.0-hr exam for a given class with a mean student mark of 70%, then set $\Delta = \pm (.10 \times 70\%) = \pm 7\%$. For this class, one would expect the mean mark for other classes on this same exam to vary somewhere between 63% and 77%. Any shortened exam with a mean inside this range would have to be considered equivalent to the 3.0-hr exam (i.e., the control group exam mean).

This empirical method of estimating delta has been used extensively in both medicine and psychology, but their standard has typically been set at $\Delta = \pm 20\%$ of the control group mean [32], [33]. From long experience teaching, we considered this value to be too large for the quantitative courses that we currently teach. For non-quantitative courses such as organizational behavior and psychology, which we have taught in the past, setting $\Delta = \pm 20\%$ of the control group mean seems reasonable given the subjectivity inherent in the marking of exams for such course material.

In the present context, a shortened final exam would be considered for all practical purposes as roughly equivalent to the full-length exam provided that average student performance on two exams did not differ by more than $\pm \Delta$ as estimated by one of these four methods. We have used all four methods and found little difference between them.

IV. Discussion

In recent years, as class sizes have grown substantially, the marking of student exams in university and college courses has become a heavy burden on the overworked professors and graduate students who must mark these exams. Even a one hour reduction in typical 3.0-hr exam length, if justified, would result in a one-third reduction in exam make-up and marking time and effort. It would be a mistake to assume that the students affected by any shortening of an exam would complain. On the contrary, provided the exam was constructed to be a fair and reasonable assessment, it seems more likely that students would greatly appreciate any reduction in stress and any extra time gained for preparing for the next exam in their busy exam schedules. Preliminary research at our university suggests this is so.

Our objective was to develop a comprehensive procedure a professor could follow to assess empirically for their own courses whether an official exam can justifiably be shortened. In Table I-Table II, we lay out a brief description of the steps we recommend instructors follow to address the question of faculty grading effort and exam length in their own courses. The primary requirement is a set of previously administered and marked exams for the target course. Faculty with an understanding of our recommended statistical procedures (correlation, reliability, and t-tests) and a statistical software package (such as SPSS or Excel) can analyze their own exam data with little time required (we estimate somewhere between 8 to 10 hours for one class of around 75 students). We recommend others consult a trained statistician (available

at most universities and colleges) for assistance. Our procedure is relatively simple and straight-forward.

TABLE I: STEPS IN OUR APPROACH FOR ASSESSING EXAM LENGTH

	Steps in Our Approach
1	To start, you must have a set of recent exams for the course in question.
2	For each student's original full-length exam, type in the marks awarded on each part of each question on the full-length exam.
3	Construct a shortened exam by selecting the subset of question parts from the full-length exam that best reflect the characteristics of a valid exam of the desired length.
4	For each exam, construct a spreadsheet algorithm that adds up the marks awarded to each student on the given exam and prints out the final mark.
5	Correlate student performance on the two exams and compute the confidence intervals for the correlation.
6	For each exam, compute the mean grade awarded and the standard deviation.
7	Use an a priori two-tailed repeated-measures t-test to test whether there is any evidence to suggest that the average grade on the shortened exam differs significantly from that for the full-length exam. If several shortened exams are to be compared with a full-length exam, then a priori Dunnett's t-tests are appropriate. Compute the confidence intervals of the difference.
8	Repeat the preceding step but using equivalence hypothesis tests and equivalence confidence intervals.
9	Carefully assess the validity, the number of questions posed, and the justifiability of test use for the shortened exam.
10	Compute coefficient alpha to estimate the reliabilities of shortened exams and the confidence intervals for alpha.

TABLE II: ADVANTAGES OF OUR APPR

No	Advantages of Our Approach
1	Minimal time and effort required to conduct the study.
2	Minimal cost to conduct the study.
3	Simple procedure that any instructor can use.
4	Permits use of a true experimental design using archival data.
5	Avoids ethical and practical problems entailed by use of conventional experimental designs.
6	Provides all the advantages of using a true experimental design for assessing cause and effect.
7	Statistical power of this design exceeds that of conventional between-subjects and within-subjects experimental designs.
8	Synthetic (or Hill-type) experimental designs permit the assessment of two additional psychometric criteria – correspondence and equivalence.
9	Our four proposed psychometric criteria complement the knowledge gained from examining the two traditional psychometric criteria – reliability and validity.
10	Our four proposed psychometric criteria are easily understood, common sense requirements compared with the frequently misunderstood classic criteria of reliability and validity.
11	Difference tests cannot, contrary to common belief, test for equivalence of group means, but equivalence tests can.
12	Equivalence testing and confidence intervals address many of the issues raised with reliance on traditional difference tests.

Preliminary results using our proposed approach suggest shortening of exams from 3.0 to 2.0 hours is frequently warranted [34]. In fact, in all 10 of the classes studied so far, shortening to 2.0 hours was justified. Moreover, these results establish generalizability across different students, lecturers, academic terms, class sizes, courses, and subject disciplines.

Readers may well question whether the results on shortening the final exam for one of our courses, say in business statistics, would apply to their own courses. Even those teaching business statistics at another university might well question whether our results are at all relevant to the construction of examinations for their own courses. Other statistics professors may emphasize different topics, construct quite different examinations, employ different teaching styles, and teach altogether different students. For courses in other subjects and disciplines, the applicability of results based on one of our courses is likely to be even more questionable. We agree. Generalization of results to other professors, students, subjects, and courses will, we believe, be highly variable and idiosyncratic. In some cases, the results will be most germane, but in others we suspect that results will be completely inapplicable. We argue, however, that this is an empirical question. We argue that any instructor questioning whether or not their own examinations can be shortened should analyze their own data to answer this question. Given the ease with which a synthetic experiment can be conducted and the marked advantages (relative to other experimental, quasi-experimental, or non-experimental designs) of such a design, there seems little reason why other professors could not empirically address the issue of exam length for their own courses.

Another objective of this paper was to re-introduce the method first used by Hill [1] to examine the question of exam length. As a variant of the traditional repeated-measures design, a Hill-type or synthetic design is a true experiment in which subjects are measured empirically in only a single experimental condition, but these observations are applied synthetically to every treatment condition. This design has been shown to meet all the criteria normally expected of any true experimental method [16]. Advantages of this design over traditional designs include substantial reductions in time and effort expended, smaller sample sizes required, lower costs, increased statistical power, and fewer threats to validity (internal, external, and statistical conclusion). Moreover, these designs are sometimes applicable in situations in which traditional designs are unwise or impractical.

Admittedly, there are limitations associated with the use of synthetic designs. The primary constraint is the limited extent to which this design can be applied. Readers interested in this approach will upon reflection realize that it cannot be used to investigate many issues in which they might be concerned. On the other hand, when they find such an application, this approach will return excellent dividends.

The crucial requirement for use of a synthetic design is independence of subjects' behaviour from experimental condition. This requirement is unusual in educational research where we are asking questions about people's behaviour. Nevertheless, when this new form of true experimental design can be used, as it can with exam length, it offers substantial advantages over traditional experimental approaches.

REFERENCES

[1] B. J. Hill, "Examination paper length: How many questions?" *Brit. J. Educ. Psych.*, vol. 48, pp. 186-195, June 1978.

[2] T. Arnone, "Large class sizes raise questions," *The Journal Queen's University*, vol. 139, no. 1, May 2011.

[3] O. Bandiera, V. Larcinese, and I. Rasul, "Heterogeneous class size effects: New evidence from a panel of students," *The Economic Journal*, vol. 120, pp. 1365-1398, Dec. 2010.

[4] K. Woodward and P. Bancroft, "Multiple choice questions not considered harmful," in *Proc. the 7th Australasian Conference on Computing Education*, Newcastle, Australia, pp. 109-116, 2005.

[5] E. Ventouras, D. Triantis, P. Tsiakas, and C. Stergiopoulos, "Comparison of examination methods based on multiple-choice questions and constructed-response questions using personal computers," *Computer and Education*, vol. 54, pp. 455-461, 2010.

[6] W. Ward and R. Bennett, *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*, New York, NY: Routledge, 2012.
[7] R. Cox, "Examinations and higher education: a survey of the literature," *Higher Education Quarterly*, vol. 21, no. 3, pp. 292-340, 1967.
[8] J. Nunnally and I. Bernstein, *Psychometric Theory* (3rd ed.), Toronto: McGraw-Hill, 1994.
[9] C. Dracup, "The reliability of marking on a psychology degree," *British Journal of Psychology*, vol. 88, pp. 691-708, 1997.
[10] J. Liu, J. Allspach, M. Feigenbaum, H. Oh, and N. Burton, *A Study of* Fatigue *Effects from the New SAT*, College Board Research Report No. 2004-5, New York, 2004.
[11] P. Ackerman and R. Kanfer, "Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions," *Journal of Experimental Psychology: Applied*, vol. 15, pp. 163-181, 2009.
[12] L. Jones and D. Thissen, "A history and overview of psychometrics," in C. R. Rao and S. Sinharay, *Handbook of Statistics*, vol. 26, pp. 1-27, Amsterdam: North Holland, 2007.
[13] D. A. Frisbie, "Reliability of scores from teacher-made tests," *Educational Measurement: Issues and Practice*, vol. 7, no. 1, pp. 25-35, 1988.
[14] G. J. Cizek, "Defining and distinguishing validity: Interpretations of score meaning and justifications of test use," *Psychological Methods*, vol. 17, pp. 31-43, 2012.
[15] G. J. Cizek, S. Rosenberg, and H. Koons, "Sources of validity evidence for educational and psychological tests," *Educational and Psychological Measurement*, vol. 68, pp. 397-412, 2008.
[16] E. S. Lee and T. Whalen, "Synthetic designs: A new form of true experimental design for use in information system development," *ACM Sigmetrics Performance Evaluation Review*, vol. 35, no. 1, pp. 191-202, 2007.
[17] W. Arthur and D. Day, "Development of a short form for the Raven Advanced Progressive Matrices Test," *Educational and Psychological Measurement*, vol. 54, no. 2, pp. 394-403, 1994.
[18] R. Hamel and V. Schmittmann, "The 20-minute version as a predictor of the Raven Advanced Progressive Matrices Test," *Educational and Psychological Measurement*, vol. 66, no. 6, pp. 1039-1046, 2006.
[19] E. S. Lee, T. Whalen, J. Sakalauskas, G. Baigent, C. Bisesar, A. McCarthy, G. Reid, and C. Wotton, "Suspect identification by facial features," *Ergonomics*, vol. 47, pp. 719-747, 2004.
[20] E. S. Lee and T. E. Whalen, "Feature approaches to suspect identification: The effect of multiple raters on system performance," *Ergonomics*, vol. 39, pp. 17-34, 1996.
[21] D. Bors and T. Stokes, "Raven's advanced progressive matrices: Norms for first-year university students and the development of a short form," *Educational and Psychological Measurement*, vol. 58, no. 3, pp. 382-398, 1998.
[22] B. Thompson, "Understanding reliability and coefficient alpha, really," in B. Thompson, Ed., *Score Reliability*, London, UK: Sage Publications, 2003, pp. 3-23.
[23] R. Henson, "Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha," *Measurement and Evaluation in Counseling and Development*, vol. 34, pp. 177-189, 2001.
[24] L. Cronbach and R. Shavelson, "My current thoughts on coefficient alpha and successor procedures," *Educational and Psychological Measurement*, vol. 64, 397-412, 2004.
[25] L. Crocker and J. Algina, *Introduction to Classical & Modern Test Theory*, Fort Worth: Harcourtolt, Brace and Jovanovich, 2008.
[26] N. Schmitt, "Uses and abuses of coefficient alpha," *Psychological Assessment*, vol. 8, pp. 350-353, 1996.
[27] X. Fan and B. Thompson, "Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial," *Educational and Psychological Measurement*, vol. 61, pp. 517-531, 2001.
[28] D. W. Zimmerman, B. D. Zumbo, and C. Lalonde, "Coefficient alpha as an estimate of test reliability under violation of two assumptions," *Educational and Psychological Measurement*, vol. 53, pp. 33-49, 1993.
[29] G. J. Cizek, D. Bowen, and K. Church, "Sources of validity evidence for educational and psychological tests: A follow-up study," *Educational and Psychological Measurement*, vol. 70, pp. 732-743, 2010.
[30] W. Schaufeli, A. Bakker, and M. Salanova, "The measurement of work engagement with a short questionnaire," *Educational and Psychological Measurement*, vol. 66, no. 4, pp. 701-716, 2006.
[31] L. Wilkinson, "Statistical methods in psychology journals," *American Psychologist*, vol. 54, pp. 594-604, 1999.
[32] L. Barker, E. Luman, M. McCauley, and S. Chu, "Assessing equivalence: An alternative to the use of difference tests for measuring disparities in vaccination coverage," *American Journal of Epidemiology*, vol. 156, pp. 1056-1061, 2002.
[33] J. Rogers, K. Howard, and J. Vessey, "Using significance tests to evaluate equivalence between two experimental groups," *Psychological Bulletin*, vol. 113, pp. 553-565, 1993.
[34] E. Lee, C. Bygrave, J. Mahar, and N. Garg, "Can exams be shortened? Using a new empirical approach to test in finance courses," *International Conference on Higher Education and Management*, November 2013.

Eric S. Lee received B.Sc. in 1972, M.Sc. in 1974, and Ph.D. in 1977 from Psychology, University of Victoria, Canada.

He was a member of the Canadian Armed Forces, after graduation, he headed a team of engineers, psychologists, and computer scientists at the Communications Research Centre in Ottawa. He has consulted with many different Canadian companies, police forces, and government departments. He has taught at Carleton University (Psychology), University of Victoria (Public Administration, and is currently at Saint Mary's University (Finance e, Information Systems, and Management Science).

Connie Bygrave received B. Comm., MBA., and Ph.D. in 2009 in management from Saint Mary's University, Halifax, Canada). She was a professor at Dalhousie University in Halifax, NS. Her research interests are employee motivation, love of the job, meaning of work, strategic alliances, ecofeminism, and spirituality in the workplace. Prior to academia, she spent over ten years in the corporate world as a marketing training specialist, a professional accountant, and a small business counselor.

Jordan Mahar received B. Comm. in 2012 from Saint Mary's University, and M.A. in 2013 in economics, from University of British Columbia, Canada. He was a member of the Mathematics Department at Dalhousie University and now works in Edmonton, Alberta.

Naina Garg is a third-year undergraduate student in economics (hons) at Saint Mary's University, Halifax, NS, Canada. She is currently working in Management Science and Management Departments as a Research Associate. She is also working on an International Development project- aimed at helping local artisan women find sustainable business opportunities in Peru. She is a member of the Commerce Society and Enactus at Saint Mary's University.