Can Exams Be Shortened? Using a New Empirical Approach to Test in Finance Courses

Eric S. Lee, Connie Bygrave, Jordan Mahar, Naina Garg, Suzanne Cottreau

Abstract**—**Marking exams is universally detested by lecturers. Final exams in many higher education courses often last 3.0 hrs. Do exams really need to be so long? Can we justifiably reduce the number of questions on them? Surprisingly few have researched these questions, arguably because of the complexity and difficulty of using traditional methods. To answer these questions empirically, we used a new approach based on three key elements: Use of an unusual variation of a true experimental design, equivalence hypothesis testing, and an expanded set of six psychometric criteria to be met by any shortened exam if it is to replace a current 3.0-hr exam (reliability, validity, justifiability, number of exam questions, correspondence, and equivalence). We compared student performance on each official 3.0-hr exam with that on five shortened exams having proportionately fewer questions (2.5, 2.0, 1.5, 1.0, and 0.5 hours) in a series of four experiments conducted in two classes in each of two finance courses (224 students in total). We found strong evidence that, in these courses, shortening of final exams to 2.0 hrs was warranted on all six psychometric criteria. Shortening these exams by one hour should result in a substantial one-third reduction in lecturer time and effort spent marking, lower student stress, and more time for students to prepare for other exams. Our approach provides a relatively simple, easy-to-use methodology that lecturers can use to examine the effect of shortening their own exams.

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

I. Introduction

Marking exams is arguably the most detested of all professorial tasks. It is a time-consuming, mind-numbing task, often requiring difficult subjective judgments even for quantitative courses. Despite almost universal revulsion for this task, few have taken the time to research the question of whether examinations can be shortened in length in order to reduce the time and effort required to mark conventional written examinations [1], [2]. By this, we refer to the style of mixed-format exam used most commonly in academe consisting of a mixture of different types of questions including problem solving (requiring detailed solutions), short essay, short answer, and even a few multiple choice questions.

Others have argued that this long-standing neglect may be attributable to the methods employed traditionally to investigate such questions [3]. We employ a new approach, based on a modification of an experimental methodology pioneered by Hill [4], which is simple to prepare and takes little time and effort to conduct.

Our primary objective was to develop a methodology that would be effective and easy enough for any instructor in higher education to use to examine the question of exam length for their own courses. A secondary purpose was to address two questions empirically: Do final exams in many higher education courses, often 3.0-hrs in length, need to be so long? Can we justifiably reduce the duration and the number of questions posed on them? We examined these questions for two different undergraduate courses in finance, for two different classes in each course, and with different students. For each class in our study, we compared student performance on six different exam versions varying systematically in length from 0.5 hours to 3.0 hours. Shortened exams were derived from the original 3.0-hr exam with proportionately fewer exam questions. Even a one-hour reduction in exam length from the traditional 3.0 hours to 2.0 hours, if warranted, should result in a one-third reduction in the total amount of marking time required. To compare shortened with full-length exams, a separate experiment was conducted on each of four classes. Because the same experimental procedures were employed with all classes, descriptions of the methods used are combined in the following sections for simplicity and brevity.

II. Method

A. Courses

Students attended one of two classes in introductory finance I (designated G and H in Table I), or one of two classes in finance II (designated I and J in Table I), both one-term third year courses presented in a series of 39 lecture-hours during a term of four months. These finance courses covered the usual mix of topics: financial analysis, working capital management, capital budgeting, the tax environment, and the role of financial intermediaries in finance I; and cost of capital, capital mix, capital and money markets, dividend policy, financial instruments and mergers, consolidations, and bankruptcy in finance II.

B. Students

A total of 224 students enrolled in the business school took one of these classes, of whom approximately 40% were male and 60% were female, most between the ages of 19 and 25.

E. S. Lee is with Finance, Information Systems, and Management Science Department, Saint Mary's University, Halifax, NS, B3H 3C3 Canada (phone: 902-420-5734; fax: 902-496-8101; e-mail: elee@smu.ca).

C. E. Bygrave is with Masters in Administrative Sciences Department, Fairleigh Dickinson University, Vancouver, BC V6B 2P6 Canada. (Canada (phone: 604-698-4494; fax: 604-682-8132; e-mail: bygrave@fdu.edu).

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

N. Garg and S. Cottreau are with Finance, Information Systems, and Management Science Department, Saint Mary's University, Halifax, NS, B3H 3C3 Canada.

C. Teacher, Examiner, Marker

All four classes were taught by the same female instructor. She made up the final exams and marked all student papers.

D. Exams

Each student wrote only a single 3.0-hour final exam in one of these courses. Each of these mixed-format exams consisted of a variety of question types: short answer, essay, problem solving, and a few multiple choice questions.

E. Experimental Design

The effect of exam length on student performance was examined using an experimental technique pioneered by Hill [4] and elaborated upon by Lee and Whalen [5], who adopted the term synthetic experimental design. Synthetic experimental designs are a form of repeated measures in which the effects of a variable are examined experimentally using synthetically generated performance scores for each student for all comparison groups except the original set of empirical performance scores. In our case, we used student performances on the 3.0-hr exam, the empirical scores, to generate the synthetic student performances for the five shortened comparison exams. For each class, there was one synthetic factor, or independent variable -- exam length.

F. Procedure

Four synthetic experiments were then conducted, one on each class, in two successive phases: an empirical phase followed by a synthetic phase [5].

1. Empirical Phase

No new data was actually collected empirically for the purposes of this study. Instead, all 224 previously graded student 3.0-hr exams were obtained from storage (university regulations dictate storage for one year), and for each student, the marks awarded for each part of a question that had been scored separately were recorded in a spreadsheet. This data constituted four empirical data sets, one for each class. Thus, a data set consisted 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).

2. Synthetic Phase

For the purposes of this study, the same original course instructor constructed five shortened versions of the 3.0-hr exam in each course (2.5, 2.0, 1.5, 1.0, and 0.5 hours). The same procedure was used to construct each shortened exam. The subset of questions on any shortened exam always constituted a subset of all the questions on the full-length exam. The subset was always selected by the instructor to produce the best possible shortened exam (i.e., appropriate for the time available) given this constraint. As a rough guide, the ratio of exam times (short/full) set the percentage of marks for questions selected from the 3.0-hr exam to be included on the shortened exam. Thus, for the 1.5-hour exam, for example, a subset of questions was selected totaling approximately (1.5/3.0 x 100 =) 50% of the original 100 marks allotted on the 3.0-hr exam.

A spreadsheet equation computed the mark a student would receive for each version of the official 3.0-hour exam. Each equation summed the marks achieved by a student on only those questions (or parts of questions) that would appear on that particular version of the 3.0-hour exam. In a few cases, the mark allotted for a particular part of a question was changed slightly to make the questions sum to the desired target.

For comparability, student marks for each shortened exam were then renormalized to a range of 0 to 100%. Thus, if the total marks on a shortened 1.5-hour exam added up to 48% (of the marks on the original 3.0-hour exam), then each student's shortened exam mark would be multiplied by 100/48 or 2.083. 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 these synthetic student performance scores was 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 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 investigation of exam length, [4] generated synthetic 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 [5].

G. Criteria for Evaluating Suitability of Shortened Exams

Our approach to assessing whether a shortened version of a traditional 3.0-hr exam could replace it as the official final exam in a course was 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. The first two criteria are traditional psychometric requirements that should be met by any examination that is used to assess student performance in a course [6], [7]. The latter four criteria are additions that we propose should also be met to justify shortening an exam. We describe each criterion briefly, 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

The performance scores of individual students on any reasonable class examination must be reliable. "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 (p. 691)" [8]. The error associated with student scores on an exam generally decreases as reliability increases. Further, "Reliability is a property of a set of test scores, not a property of the test itself (p. 25)" [2].

We estimated internal-consistency reliability because it can be assessed from a single administration of a test and it is the most frequently reported measure of reliability [9]. Following common psychometric practice, we estimated internal consistency reliability using Cronbach's 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 [6], [10]. "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 (p. 28)" [2].

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? In general, "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 (p. 353)" [11]. Nevertheless, the higher the reliability, the better it is for making decisions about students.

However, a standard for reliability should always be specified and a rationale provided for that value [12]. We argue that alpha should equal or exceed roughly 0.80 ($\alpha \geq 0.80$) for the finance courses examined in our study, with the caveat that even higher reliabilities are preferable. Our rationale for this criterion is based on three considerations. First, reliability for quantitative courses is undoubtedly generally higher than that for non-quantitative courses [8]. Second, we note that reliability for a final exam in these finance courses need not be so high since final grades in our institution are based not just on student performance on the final exam, but also on midterms, assignments, projects, and other work which can greatly increase overall reliability in a course [2]. Third, decisions affecting students are typically based on their performance in many different courses. Reliabilities based on a collection of courses can easily exceed 0.90 even when the reliability of any single course is much lower [8]. Nevertheless, to be on the conservative side, we set the minimum standard to be exceeded for reliability of exam scores in a given class of these finance courses as $\alpha \geq 0.80$.

2. Validity Criterion

Validity, or the degree to which exams measure what they purport to measure, is arguably the most important criterion to be satisfied when considering alternative exams. Claims and decisions based on shortened exams should, therefore, be as valid as those based on current full-length 3.0-hr exams.

Internal consistency reliability does not assure validity [13]. It does, however, set an upper bound on the possible validity associated with an exam. Consequently, we examine reliabilities associated with exam scores to assess evidence for validity based on internal structure.

We believe that two other sources of validity evidence for the interpretation of student performance on exams should be examined: face validity evidence and evidence based on content validity [14]. 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 [14].

Face validity refers to the degree to which the course instructor subjectively judges an exam to be fair, reasonable, and appropriate, that is, how well the exam covers the knowledge and skills taught in the course. We asked whether or not she would be willing to use each of the six exam versions as the official final exam in her courses. As well, we asked her 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.

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 effective planning and design [6]. This approach was used for construction of all 3.0-hr exams but was inadmissible for the shortened versions. For shortened exams, therefore, we relied on assessing content validity using a short questionnaire. We asked the instructor two questions (using a 5-point Likert scale of 1 = not very well to 5 = very well). First, how well did each of the six exams cover all important topics covered 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).

3. Justification-of-Use Criterion

Traditionally, reliability and validity have been the only psychometric properties examined when developing tests. However, reliability and validity are frequently misunderstood and misinterpreted by both instructors and researchers [2]. We argue that four other informal psychometric properties – justification of test use, number of exam questions, equivalence, and correspondence – provide additional, more easily understood, insight into the assessment of the suitability of shortened exams as replacements for current 3.0-hr exams. While these four criteria do not provide additional independent sources of evidence for or against shortening (just as reliability and validity are not independent), they are understood and interpreted correctly more easily.

Cizek's justification of test use is important "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 (p. 741)" [15]. We believe that four sources of justification can be examined for assessing this criterion in our studies [15], [16]. 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 replacing the current official exam with a shortened exam?

4. Number of Exam Questions Criterion

The number of questions (separately scorable), or parts of questions, on an exam is directly related to both reliability and validity. The Spearman-Brown prophecy formula (derivable from classical test theory) indicates how reliability (in our case, as estimated by coefficient alpha) can be increased by simply increasing the number of questions on an exam [6]. 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.

A second advantage centers on our caution that a minimum of 10 questions should appear on any final exam to ensure adequate reliability [6]. The investigation by [4] shows the value of this rule. If Hill had followed this rule, he would have realized that none of his 3.0-hr engineering exams be justifiably shortened since each of his exams had only 5-6 questions.

5. Correspondence Criterion

A third common-sense criterion that should, in our opinion, be met by any suitable shortened (replacement) exam is that students should perform as well on such an exam as on the official 3.0-hr exam. That is, we expect student marks on a suitable shortened exam to correlate highly with that on the 3.0-hr exam. It is analogous to the criterion employed by researchers developing shortened forms of already established full-length psychological tests [17].

6. Equivalence (and Difference) Criterion

A fourth common-sense criterion is that student performance on a suitably shortened exam should, on average, be roughly equivalent to that on the current 3.0-hr exam. Most professors recognize that average student performance in a course will vary somewhat from one class to another. Given that exams are never the same from one class to another, the students differ, and how one teaches varies over time. Nevertheless, most would agree that class averages, while never exactly the same, should be roughly equivalent. For a shortened exam to be considered an acceptable substitute for the 3.0-hr exam, we hypothesized, first, that average student performance on that exam would not deviate significantly from that on the 3.0-hr exam, and second, that average student performance on the two exams should be roughly equivalent.

H.Statistical Analysis

In assessing the degree to which each shortened exam met the selection criteria described earlier, we considered confidence intervals, effect sizes, nil null hypothesis testing, testing of assumptions, reliabilities, correlations, and the results of traditional difference significance tests and the newer equivalence tests. We discuss each, providing a rationale and detailing precisely what we did and the nature of the statistical analyses used. We address many of the criticisms that have been made of traditional statistical hypothesis tests and their misinterpretation [18].

1. Confidence Intervals

Following the recommendations of [12] and [19], we report confidence intervals for all analyses: for reliabilities, for correlations, and for differences and for equivalences between mean student performances on shortened versus 3.0-hr exams. Confidence intervals offer many advantages over reliance on only traditional statistical hypothesis tests. First, p values used in hypothesis tests are confounded measures of study effect sizes, and therefore, can be seriously misleading. Second, they provide easily understood and readily interpreted estimates [20]. Third, confidence intervals have a close connection with statistical testing: “Noting that an interval excludes a value is equivalent to rejecting a hypothesis that asserts that value as true (p. 534)” [20]. Later, we discuss in detail precisely which confidence intervals we used.

2. Effect Sizes

Effect sizes assess practical, as opposed to statistical, significance. They measure the magnitude of an effect [21]. For all correlational and reliability studies, effect size estimates are given directly by the reported correlations or reliabilities (or by squaring them, r^2). For the analyses of differences or equivalences between means (of student scores on short and full-length exams), we report generalized eta squared values which directly estimate the percentage of variation explained [22], [23].

3. Nil Null Hypothesis Testing

We follow the advice of Thompson [19] to avoid the use of nil null hypothesis testing in which researchers test a null hypothesis of no difference. Instead, we test whether a precise criterion value is exceeded or not. The problem with use of nil null hypothesis testing is that any such null hypothesis can always be rejected by simply testing a large enough sample.

4. Testing of Assumptions

Both difference and equivalence tests are based on the assumptions of normality and equality of variances for each pairwise comparison of a shortened versus full-length 3.0-hr exams. However, provided that one has equal sample sizes, as we did, “there is still reason to believe that normality is not a crucial assumption and that the homogeneity of variance assumption can be violated without terrible consequences. (p. 340)” Howell [24] advises against performing direct statistical tests of homogeneity (such as Levene’s test and the Fmax test); he notes, however, that any ratio of largest to smallest variances in these pairwise comparisons greater than 4.0 would be a cause for concern. (There were none.)

5. Correspondence Analyses

A student who does well on a 3.0-hr exam should do equally well on a shortened exam version to be considered an acceptable substitute. This hypothesis was tested in each experiment using a priori linear correlations (Pearson r) between student performance on the control exam and that on each shortened exam. We expected that correlations should exceed a minimum of $r \geq 0.90$ so that at least roughly 80% (r^2) of the total variation in student performance on the full-length exam could be explained by the variation in student performance on the shortened exam. Confidence intervals for each correlation were computed using the procedure described by Fan and Thompson [12] and the “R2” computer program of

Steiger and Fouladi [25], [26]. Noncentral, as opposed to conventional central approaches, are appropriate for computing confidence intervals for correlation estimates when (a) effect sizes are large and (b) sample sizes are small.

6. Reliability Analyses

We followed the advice of Fan and Thompson to compute coefficient alpha and reporting confidence intervals for reliability estimates [12]. Noncentral, as opposed to conventional central approaches, are most appropriate for computing confidence intervals for reliability estimates when (a) effect sizes are large and (b) sample sizes are small. In this study, we expected some small sample sizes and few questions on some exams. Since we wish to generalize the results of our studies to other students and other exams, we controlled for random effects using the computer program "R2" to estimate our reliability confidence intervals for coefficient alpha for each exam [12], [25], [26].

7. Equivalence and Difference Analyses

These analyses focused on comparing mean student performance on the 3.0-hr exam with that on each shortened exam. For a shortened exam to be considered an acceptable substitute, we hypothesized first that average student performance on that exam would not deviate significantly from that on the 3.0-hr exam, and second, that average student performance on the two exams should be roughly equivalent. To test the first part of this hypothesis, we used traditional statistical tests, sometimes referred to as difference hypothesis testing. However, in traditional hypothesis testing, a null hypothesis conclusion does not offer proof that the null hypothesis of no difference (or equivalence) is correct. Rather, such a conclusion only permits one to conclude that there is insufficient evidence to reject the null hypothesis [18].

The second part of this hypothesis asserts that, for a shortened exam to be considered an acceptable substitute for the full-length exam, average student performance on that exam should be statistically equivalent to that on the full-length exam. This hypothesis must be tested using equivalence testing, [27], [28]. In equivalence testing, a null hypothesis conclusion only permits one to conclude that there is insufficient evidence to reject the equivalence null hypothesis.

Therefore, to compare average student performance on a given full-length 3.0-hour exam with that on each shortened version of the same exam, we conducted two complementary sets of statistical analyses using simple conventional repeated-measures t-tests and confidence intervals plus equivalence repeated-measures confidence intervals and t-tests. We follow the lead of those advocating the use of both difference testing and equivalence testing to examine the same data [27], [28].

Both aspects of our hypothesis (difference and equivalence) involve the a priori (or planned) comparison of a control group (the 3.0-hour exam) with several treatment groups (the five shortened exam versions). Consequently, to test both aspects of this hypothesis, Dunnett's t-test for a priori or planned multiple comparisons with repeated measures data is appropriate with the 3.0-hour exam serving as the control group and the five shortened exams as the treatment groups [29], [30], [24]. We used traditional repeated-measures t-tests to compare mean performance on the full-length exam for a class with mean performance on each shortened version of that exam but assessed significance using Dunnett's special t-tables [29], [30], [24]. This test is more powerful than any other tests that aim at holding the familywise error rate at or below α. Dunnett's tests were used without any preliminary overall F-test, as advocated by many [24], [32]. The family wise error rate was set at $\alpha = 5\%$. For difference tests, we used two-tailed Dunnett's t-values to assess significance and to compute confidence intervals. Note, however, that for equivalence tests, one-tailed Dunnett's t-values must be used to assess significance and to compute confidence intervals. We followed the advice of Howell [31] and Maxwell [33] against using a pooled error term based on all the data and instead used "only the data involved in those contrasts to run the contrasts (p. 13)" [31]. The only comparisons of interest in each experiment were between the control exam and each shortened exam. Comparisons between pairs of the five shortened versions of each exam were not of interest and so were not tested.

When using both difference and equivalence tests to compare average student performance on a full-length exam with that on any shortened version, four outcomes are possible: both difference and equivalence tests are significant, equivalence test only is significant, difference test only is significant, and neither test is significant [27], [28]. If both tests are significant, then we will conclude that, while we have solid evidence that a real difference exists, the difference is so trivial (given our finding of equivalence) that the two exams are, for all practical purposes, equivalent. If only the equivalence test is significant, then we will conclude that we have true equivalence (at least within the tolerance of our tests). Of course, failure to find evidence of a significant difference using difference testing cannot be interpreted as evidence of equivalence. If only the difference test is significant, then we will conclude that we have evidence of a difference in average student performance between the two exams (note that a failure to find equivalence by the equivalence test does not permit us to conclude that there is evidence of a real difference). If neither test is significant, then we must defer our decision as we have insufficient evidence of either difference or equivalence.

I. *Estimating Variability in Mean Grades on Final Exams*

Equivalence tests require the estimation of delta (Δ) which determines the range (±Δ) within which the observed difference between mean student performances between shortened and full-length exams could normally be expected to fall [27], [28]. In both medicine and psychology, the standard has typically been set at $\Delta = \pm 20\%$ of the control group mean. From long experience teaching, we considered this value to be too large. In our opinion, we would reject as being non-equivalent any shortened exam that exceeded roughly ±10% of the control group mean for each class (for the 3.0-hr exam). Thus, if the 3.0-hr exam for a given class has a mean student mark of 70%, then we set Δ = ± (0.10 x 70%) = ± 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.

III. Results

All results except for justifiability are displayed in Table I. Because all statistical results are displayed there, these values are not repeated in the following sections. Justifiability results are discussed for each exam in the following sections. We detected no evidence of serious violation of assumptions.

A. *Finance I*

The two classes (G and H) in Finance I provide independent sources of empirical evidence on the course.

1. 3.0-hr Exam

Full-length 3.0-hr exams can be evaluated on just four of our six psychometric selection criteria: validity, justifiability, number of exam questions, and reliability. The instructor judged both face and content validity of this exam to be acceptable (≥3 on 5-point Likert rating scales) on all dimensions that we measured. Justifiability of test use of this long 3.0-hr exam, however, is low given that the students and instructor currently invest the historical maximum amount of time and effort either to write or to mark such a lengthy exam. The number of questions posed on this exam in the two classes was 37, almost four times our minimum requirement of 10. Reliability was not a problem given that the lower bound of the 95% confidence interval for α exceeded .84 in both classes.

2. 2.5-hr Exam

The 2.5-hr shortened exam met all six criteria for both classes G and H. The instructor somewhat surprisingly judged this shortened exam to be superior to the official 3.0-hr exam on both face and content validity. Internal structure validity, based on estimates of reliability, was also very high (α > 0.84 for class G and 0.87 for class H). Justifiability of shortened test use was also high, given that the exam would take approximately 16-17% less time and effort to make up and grade than the current official exam. The shortened 2.5-hr exam consisted of approximately 30 separately scorable questions, well above the minimum standard of 10.

Reliability clearly did not differ between the 2.5-hr and 3.0-hr exams in either class. As well, the lower bound of the reliability confidence interval for coefficient alpha in each class for the 2.5-hr exam far exceeds the target that we set.

The correspondence between student performance on the 2.5-hr and 3.0-hr exams was very, very high in both classes G and H. The lower bound of the 95% confidence interval for r was virtually identical for the two classes, strongly suggesting that the population correlation exceeds 0.98. Clearly, shortening the official current exam by a half hour introduces little error ($1 - r^2 = 4.0\%$), and students in each class had virtually the same ranking on the two exams (short and long).

For class H, mean student performance on this shortened exam was statistically equivalent to that on the control-group 3.0-hr exam. There was no evidence of a significant difference between the means on these two exams. For class G, though mean student performance on the 2.5-hr exam differed significantly from that on the 3.0-hr exam, this difference must be considered trivial because mean performances on these two exams are equivalent. The magnitude of this equivalence effect was substantial at over 78% of the total variation in 3.0-hr exam scores explained by variation in student performance on the 2.5-hr exam and contrasts strongly with the small effect sizes observed for the difference effect in this class (15%).

3. 2.0-hr Exam

The 2.0-hr shortened exam met almost all six selection criteria in the two classes. This exam contained approximately 23 questions, well above the minimum standard set of 10. Furthermore, the number of questions posed on this exam was so large that it suggests our reliability criterion should be met. The course instructor judged the evidence for content validity of this shortened exam to be equivalent to that on the 3.0-hr exam. The high reliability of exam scores, discussed below, suggests internal structure validity is high. The instructor made no judgment of face validity using the 5-point Likert scale, thereby making it impossible to compare face validity on this shortened exam with that on the 3.0-hr exam. In terms of the overall judgment of face validity, the instructor judged the shortened exam to be unacceptable. Justifiability of using this shortened test was also high, given that time and labour required to prepare and mark a final exam of this length would be 33% less than that required by a 3.0-hr exam.

The correspondence between student performance on the 2.0-hr and 3.0-hr exams, as estimated by the lower bound of the 95% confidence interval for r, was high and virtually identical for classes G and H, 0.95 and 0.96. Students had almost identical rank orderings on the 2.0-hr and 3.0-hr exams with little error (9.8%).

The reliability of exam scores did not differ between the 2.0-hr and 3.0-hr exams in either class. The lower bound of the reliability confidence interval for alpha for both classes exceeded the target that we set as standard (α ≥ 0.80).

For both classes G and H, mean student performance on the shortened 2.0-hr exam was statistically equivalent to that on the full-length 3.0-hr exam. Though the two exams did differ significantly for class G, this difference must be considered trivial because effect size is small and the means of the two exams are statistically equivalent. The two exams did not differ significantly for class H.

4. 1.5-hr Exam

The 1.5-hr exam met most selection criteria in classes G and H. This shortened exam contained just 15 questions. While this is above the minimum standard we set, this relatively small number of questions might prove problematic for both reliability and validity. In fact, the evidence for face validity was equivocal. The course instructor judged the evidence for face validity of this shortened exam to be inadequate though she also rated this exam as marginal or moderate using the 5-point Likert scale. Content validity was also judged to be moderate or nearly acceptable on both coverage and proportional allocation of marks. There is marginal or moderate evidence of internal structure validity as well given that reliability is only moderately high. Justifiability of test use for this shortened exam is high since a 50% reduction in official exam length should result in a marked 50% reduction in time and effort spent marking.

The correspondence between student performance on the 1.5-hr and 3.0-hr exams was nearly acceptable for classes G and H with the lower bound of the 95% confidence interval for r falling just below or just above our target. Rank orderings of students on exam performance differed somewhat between the two exams with a relatively high degree of error (19%).

The reliability of exam scores on the 1.5-hr exam was significantly below that for the current official 3.0-hr exam (the two 95% confidence intervals in each class did not overlap). The shortened exam scores for both classes had reliabilities that failed to meet our target ($\alpha \geq 0.80$).

For both classes G and H, mean student performance on the shortened 1.5-hr exam was statistically equivalent to that on the 3.0-hr exam. Mean performances on the two exams did not differ significantly in either class.

5. 1.0-hr and 0.5-hr Exams

Both of these exams failed to meet most of our selection criteria in both classes and will not be discussed further here (see Table I for details).

B. Finance II

The results for two synthetic experiments, one conducted on each of two classes, I and J, in Finance II are discussed.

1. 3.0-hr Exam

Full-length 3.0-hr exams can be evaluated on just four of our six psychometric selection criteria: number of exam questions, reliability, validity, and justifiability. The number of questions was so large (47), almost five times our minimum requirement, that reliability was unlikely to be a problem. In fact, reliability of this exam, as estimated by the 95% confidence interval for α, far exceeded for both classes the minimum standard we had set of $\alpha \geq 0.80$. The instructor judged both face and content validity of this exam to be the highest possible on all dimensions that we measured. Justifiability of test use of this long 3.0-hr exam, however, is low given that the students and instructor currently invest the historical maximum amount of time and effort either to write or to mark such a lengthy exam.

2. 2.5-hr Exam

The 2.5-hr shortened exam met all six criteria for both classes I and J. This particular exam contained 36 separately scorable questions, over three times as many as we set as the minimum standard of 10. Such a high number of questions strongly suggests this shortened exam exceeds our objective for a suitable shortened exam. The course instructor judged face and content validity to be very good. Given the high reliability in each class, evidence for internal structure validity was very high ($\alpha > 0.82$). Justifiability of shortened test use was reasonable and higher than that for the current 3.0-hr exam given that student writing time and instructor grading time would be reduced by roughly 17%.

The correspondence between student performances on the 2.5-hr and 3.0-hr exams was very high in both classes I and J. The lower bound of the 95% confidence interval for *r*, strongly suggests that the population correlation exceeds 0.98, far exceeding our minimum standard. Such a high correlation indicates that there is little error (4.0%) introduced by replacing the current exam with an exam one-half hour shorter in length. Students in each class had virtually the same ranking on both this shortened exam and the current 3.0-hr exam.

The reliability of student scores on this exam did not differ from that for the current 3.0-hr exam in either class I or J. As well, the lower bound of the reliability confidence interval in each class for the 2.5-hr exam exceeded our target of 0.80.

Correspondence, as measured by the correlation between student performance on the 2.0-hr and 3.0-hr exams, was very high for both classes, at 0.95 to 0.98 and 0.96 to 0.98. These 95% confidence interval estimates strongly suggest that the population correlation between student scores on the shortened exam and the 3.0-hr exam exceeds 0.95. Students who did well on the official 3.0-hr exam did equally well on the shortened 2.0-hr exam. Furthermore, error was small (9.8%).

The reliability of exam scores did not differ between the 2.0-hr and 3.0-hr exams in class I although it was significantly smaller in class J. However, the lower bound of the reliability confidence interval for alpha for both classes exceeded the target that we set as standard ($\alpha \geq 0.80$).

TABLE I
COMPARISON OF STUDENT PERFORMANCE ON FINAL EXAMS OF VARIOUS LENGTHS IN 4 FINANCE CLASSES (MEAN DIFFERENCES, MEAN EQUIVALENCES, CORRELATIONS, RELIABILITIES, AND VALIDITIES)

	Exam length		Student exam performance (%)		Analyses of criteria												
					Mean difference[a]			Mean equivalence[b]		Correspondence		Reliability		Validity			
Class	t	k	M (SD)	MD (SE)	t_{diff} (η^2_G)	95%CI$_{diff}$	±Δ	t_{equiv} (η^2_G)	90%CI$_{equiv}$	r(t,3.0)	95%CI$_r$	α	95%CI$_\alpha$	F1	F2	C1	C2
G	3.0	37	63.3 (16.9)									**0.85**	0.84, 0.86	Y	3	3	3
n=60	2.5	30	62.2 (17.6)	1.16 (0.36)	**3.21** (15)	0.2, 2.1	±6.3	**-14.28** (78)	0.3, 2.0	**0.99**	0.98, 0.99	**0.86**	0.84, 0.87	Y	4	4	4
	2.0	23	61.7 (18.9)	1.57 (0.60)	**2.61** (10)	0.0, 3.1	±6.3	**-7.88** (51)	0.2, 3.0	**0.97**	0.95, 0.98	**0.87**	0.85, 0.87	N	**?**	3	3
	1.5	15	64.9 (16.5)	-1.56(0.77)	-2.04 (7)	-3.6, 0.4	±6.3	**6.16** (39)	-3.3, 0.2	**0.94**	0.90, 0.96	0.73	0.66, 0.73	N	3	3	3
	1.0	12	56.4 (19.5)	6.95 (1.09)	**6.39** (40)	4.1, 9.8	±6.3	0.60 (0)	4.5, 9.4	0.90	0.84, 0.94	0.73	0.66, 0.74	N	3	3	3
	0.5	9	69.3 (15.4)	-5.97(1.29)	**4.63** (27)	-9.3,-2.6	±6.3	0.26 (0)	-8.9,-3.0	0.81	0.70, 0.88	0.46	0.23, 0.42	N	1	1	1
H	3.0	37	67.2 (19.2)									**0.86**	0.85, 0.87	Y	3	3	3
n=57	2.5	30	66.8 (18.8)	0.42 (0.41)	1.02 (2)	-0.6, 1.5	±6.7	**-15.32** (81)	-0.5, 1.4	**0.99**	0.98, 0.99	**0.88**	0.87, 0.89	Y	4	4	4
	2.0	23	66.5 (20.0)	0.75 (0.60)	1.25 (3)	-0.8, 2.3	±6.7	**-9.92** (64)	-0.6, 2.1	**0.97**	0.96, 0.98	**0.88**	0.86, 0.89	N	**?**	3	3
	1.5	15	68.0 (18.4)	-0.81 (0.79)	-1.03 (2)	-2.8, 1.2	±6.7	**7.46** (50)	-2.6, 1.0	**0.95**	0.92, 0.97	0.79	0.74, 0.80	N	3	3	3
	1.0	12	61.7 (20.4)	5.54 (0.96)	**5.80** (38)	3.1, 8.0	±6.7	-1.21 (3)	3.4, 7.7	0.94	0.89, 0.96	0.72	0.64, 0.73	N	3	3	3
	0.5	9	71.1 (19.3)	-3.91 (1.11)	**-3.51** (18)	-6.8,-1.0	±6.7	**2.51** (10)	-6.4,-1.4	0.90	0.84, 0.94	0.70	0.59, 0.70	N	1	1	1
I	3.0	47	67.7 (14.9)									**0.85**	0.83, 0.85	Y	5	5	5
n=52	2.5	36	62.0 (15.3)	5.70 (0.32)	**17.57**(86)	4.9, 6.5	±6.8	**-3.44** (19)	5.0, 6.4	**0.99**	0.98, 0.99	**0.84**	0.82, 0.85	Y	4	4	4
	2.0	29	64.7 (16.2)	3.02 (0.54)	**5.56** (38)	1.6, 4.4	±6.8	**-7.00** (49)	1.8, 4.2	**0.97**	0.95, 0.98	**0.82**	0.80, 0.83	N	3	4	4
	1.5	25	66.6 (14.7)	1.14 (0.79)	1.45 (4)	-0.9, 3.2	±6.8	**-7.16** (50)	-0.7, 3.0	0.92	0.87, 0.96	0.74	0.70, 0.76	N	2	3	3
	1.0	17	63.0 (13.4)	4.76 (1.09)	**4.37** (27)	1.9, 7.6	±6.8	-1.87 (6)	2.3, 7.2	0.85	0.75, 0.91	0.53	0.43, 0.54	N	1	2	1
	0.5	10	60.0 (16.5)	7.66 (2.24)	**3.42** (19)	1.9, 1.3	±6.8	0.38 (0)	2.5, 12.8	0.48	0.23, 0.66	0.36	0, 0.29	N	1	1	1
J	3.0	47	67.9 (17.2)									**0.88**	0.87, 0.89	Y	5	5	5
n=55	2.5	36	62.5 (16.8)	5.46 (0.33)	**16.56**(84)	4.6, 6.3	±6.8	**-4.06** (23)	4.7, 6.2	**0.99**	0.98, 0.99	**0.86**	0.85, 0.87	Y	4	4	4
	2.0	29	64.5 (17.3)	3.40 (0.52)	**6.54** (44)	2.1, 4.7	±6.8	**-6.54** (44)	2.2, 4.6	**0.98**	0.96, 0.98	**0.84**	0.82, 0.85	N	3	4	4
	1.5	25	67.0 (15.6)	0.90 (0.80)	1.13 (2)	-1.2, 3.0	±6.8	**-7.38** (50)	-0.9, 2.7	0.94	0.89, 0.96	0.78	0.75, 0.79	N	2	3	3
	1.0	17	63.4 (18.5)	4.50 (1.26)	**3.56** (19)	1.2, 7.8	±6.8	-1.83 (6)	1.6, 7.4	0.86	0.77, 0.92	0.76	0.71, 0.77	N	1	2	1
	0.5	10	63.9 (18.8)	3.98 (2.31)	1.72 (5)	-2.0, 9.9	±6.8	-1.22 (3)	-1.3, 9.2	0.55	0.33, 0.71	0.52	0.36, 0.51	N	1	1	1

Note. Bold numbers indicate significance ($p < 5\%$); statistics classes = A – F; finance I classes = G – H; finance II classes = I – J; t = time allotted for students to complete exam; k = number of exam questions; MD = mean difference between shortened and full-length (3.0-hr) exam; SE = standard error of the difference between short and 3.0-hr exam means; t_{diff} = difference hypothesis testing (traditional) t-test value for within-subjects comparison of means of short and 3.0-hr exams; t_{equiv} = equivalence hypothesis testing t-test value for within-subjects comparison of means of short and 3.0-hr exams; η^2_G = generalized eta squared effect size estimate = the percentage of total variation explained; r(short, 3.0) = Pearson *r* correlation (r's significantly > 0.70 are boldfaced) between student performance on the 3.0-hour exam and that on each shortened version of the same exam (2.5, 2.0, 1.5, 1.0, and .5 hours); α = reliability estimated by Cronbach's coefficient alpha based on scores for each exam (α's significantly > 0.70 are boldfaced); judged validity = acceptability of exam as official course exam as judged by examiner (Yes or No); rated validity = acceptability of exam as official course exam as judged by examiner (on a 5-pt Likert scale: not at all acceptable 1 2 3 4 5 very acceptable).

[a]Difference testing with conventional paired-samples *t*-test results reported but significance assessed using Dunnett's (1955, 1964) tabled two-sided critical *t* values (2.58, 2.58, 2.59, and 2.58 for classes A to J, respectively; α = 5%; 2-sided; and k = 6 groups) which are appropriate for comparing each shortened exam with a control (the 3.0-hour exam). Significant differences are in boldface.

[b]Equivalence testing comprised two one-sided paired-samples *t*-tests, but setting μ_d = the equivalence criterion of ±Δ = ±10% of control-group (3.0-hr exam) mean and with significance assessed using Dunnett's (1955) one-sided critical *t* values (2.28, 2.28, 2.29, and 2.28 for classes A to J, respectively; α = ±5% for each one-sided test; and k = 6 groups). Only the smaller of the two *t*-tests (which also has the larger *p* value) is reported. Significant equivalencies are printed in boldface.

Open Science Index, Educational and Pedagogical Sciences Vol:8, No:1, 2014 publications.waset.org/9997150.pdf

For both classes I and J, mean student performance on these two exams (the 2.0-hr and 3.0-hr exams) were both statistically equivalent, and statistically different. Though the two exams did differ significantly for both classes, these differences must be considered trivial because the means of the two exams in each class are statistically equivalent.

3. 1.5-hr Exam

The 1.5-hr exam met most selection criteria in classes I and J. This shortened exam contained just 15 questions. While this is above the minimum standard we set of 10, this relatively small number of questions might prove problematic for both reliability and validity. In fact, the evidence for face validity was equivocal at best. The course instructor judged the evidence for face validity of this shortened exam to be inadequate though she also rated this exam as marginal or moderate using the 5-point Likert scale. Content validity was also judged to be moderate or marginally acceptable on both coverage and proportional allocation of marks. There is marginal or moderate evidence of internal structure validity as well given that reliability is only moderately high ($\alpha > 0.70$). Justifiability of test use for this shortened exam is high since a 50% reduction in official exam length should result in a marked 50% reduction in time and effort spent marking.

The reliability of exam scores on the 1.5-hr exam was significantly below that for the current official 3.0-hr exam (the two 95% confidence intervals in each class did not overlap). The shortened exam scores for both classes I and J had lower bound confidence interval reliabilities of 0.70 and 0.78 that failed to meet our target 0.80.

The correspondence between student performance on the 1.5-hr and 3.0-hr exams was marginal for classes I and J with the lower bound of the 95% confidence interval for r falling just below our target. Rank orderings of students on exam performance differed somewhat between the two exams with a relatively high degree of error (24.3%).

For both classes I and J, mean student performance on the shortened 1.5-hr exam was statistically equivalent to that on the 3.0-hr exam. Mean performances on the two exams did not differ significantly in either class.

4. 1.0-hr and 0.5-hr Exams

Both of these shortened exams failed to meet most of our selection criteria in both classes and will not be discussed further here (see Table I for details).

IV. Discussion

Our primary objective was to develop a methodology that any lecturer could use to assess whether or not a long 3.0-hr exam could be justifiably shortened. We argue that the method that we propose in this paper achieves this goal. First, our proposed method does not require complex, expensive, or time-consuming research. Instead, our approach permits the use of archival data (already marked official student 3.0-hr exams) with not much additional time required. Most professors retain graded exams for a year or more. To determine whether such a long official exam can be justifiably shortened in the future, the lecturer need complete only four tasks. First, only a single shortened exam of the desired length needs to be constructed by choosing a subset of the questions on the full-length official exam. (In our study, we constructed for research purposes five different shortened versions of the 3.0-hr exams. This is unnecessary generally.) For example, a 2.0-hr shortened exam can be designed by selecting questions which sum to two-thirds of the marks on the 3.0-hr exam. Second, the marks awarded to each student in a class for all parts of all questions that were assigned a separate grade on the full-length exam must be typed into a spreadsheet. Third, a spreadsheet equation must be designed to compute the mark achieved by each student in a class on both the full-length and the shortened exams. Fourth, student performance on the two exams must be compared on the six selection criteria described in method section G. None of these tasks require much time or effort (when compared with conventional testing methods). If a shortened exam is warranted, then the time invested in this assessment procedure could be recovered in the time saved in subsequent marking of shorter exams.

A secondary purpose of this paper was to use our proposed approach to demonstrate to lecturers that shorter tests under some conditions are satisfactory. We assessed whether the current official 3.0-hr final exams in two university finance courses could justifiably be shortened, and if so, by how much. Based on our empirical comparison of shortened and full-length exams in these two courses on six selection criteria, we found the 3.0-hr exams could arguably be shortened to 2.0 hours without materially affecting student performance. This conclusion was confirmed in two separate, independent assessments of 3.0, 2.5, 2.0, 1.5, 1.0, 0.5-hr exams in each course (as we tested two classes in each course). The 2.0-hr exam in each course met all six of our proposed selection criteria: reliability, validity, justifiability, correspondence, equivalence, and number of questions posed on an exam. The 2.0-hr (and 2.5-hr) exams were as reliable as the 3.0-hr exams; reliability of the 2.0-hr exams exceeded our standard for coefficient alpha of 0.80; student performance on the shortened 2.0-hr exams correlated highly with that on the 3.0-hr exams in each class (the lower bound of the 95% confidence interval for this correlation exceeded 0.95 in all four classes); mean student performances on the shortened 2.0-hr exams were statistically equivalent to that on the 3.0-hr exams in each class; the number of separately scorable questions on each exam was two to three times the minimum standard we had set ($k \geq 10$) and more than high enough to suggest that reliability and validity would not be adversely affected by shortening; evidence for face, content, and internal structure validity was satisfactory for shortened 2.0-hr exams and in some cases was even better than that for the 3.0-hr exam (though borderline in one other class); and finally, shortened test use is justified given that over 33% of the time and effort expended by lecturers making up and marking final exams in these two finance courses could be saved by shortening the current exam length from the three to two hours. Moreover, students would welcome the reduction in stress and extra time gained for preparing for subsequent exams if exam length were shortened.

We suspect that other lecturers who teach higher education courses in finance and other fields might also be able to reduce their current final exam lengths below 3.0 hrs. This is, however, an empirical question. Readers may well question whether the results on shortening the final exam for one of our courses in finance would apply to their own courses. Even those teaching finance at another university might well question whether our results are at all relevant to the construction of examinations for their own courses. Other finance professors may emphasize different topics, construct different examinations, employ different teaching styles, and teach 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 our results to other professors, students, subjects, and courses will, we believe, be somewhat variable and idiosyncratic. In some cases, the results will be most germane, but in others we suspect that our results will be completely inapplicable. Further testing should answer this question.

A limitation of our approach is that it is predicated on the assumption that student answers to a question on a shortened exam would not differ from that given for the same questions posed on the 3.0-hr exams. Further research is needed to assess the impact that exam length has on students' cognitive fatigue, effort, and performance [34]. We argue that there is little reason to expect student answers for a question to differ between different versions of an exam if the same amount of time was given to answer the question on both exams. Another limitation is that our study was confined to testing only one professor and two courses in a single subject area. However, recent research in our lab has produced very similar results for a variety of students, professors, courses, and subject areas. Some instructors may be unfamiliar with our recommended statistical techniques while others may view them as complex. However, most universities employ applied statisticians who can help.

References

[1] R. Cox, "Examinations and higher education: a survey of the literature," in *Higher Education Quarterly*, 21(3):pp. 292-340, 1967.

[2] D. A. Frisbie, "Reliability of scores from teacher-made tests," in *Educational Measurement: Issues and Practice,* 7, no.1: 25-35, 1988.

[3] E. Lee, C. Bygrave, J. Mahar, and N. Garg, "Can higher education exams be shortened? A proposed methodology for studying this issue," in *International Conference on Higher Education and Management*, submitted for publication, October 2013.

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

[5] E. S. Lee, and T. Whalen, "Synthetic designs: A new form of true experimental design for use in information system development," in *ACM Sigmetrics Performance Evaluation Review*, 35(1), pp. 191-202, 2007.

[6] J. Nunnally, and I. Bernstein, in *Psychometric theory* (3rd ed.). Toronto: McGraw-Hill, 1994.

[7] L. Crocker, and J. Algina, *Introduction to classical & modern test theory*. Fort Worth: Harcourtolt, Brace and Jovanovich, 2008.

[8] C. Dracup, "The reliability of marking on a psychology degree," in *British Journal of Psychology*, 88, pp. 691-708, 1997.

[9] B. Thompson, "Understanding reliability and coefficient alpha, really," in B. Thompson (Ed.), *Score reliability* (pp. 3-23). London, UK: Sage Publications, 2003.

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

[11] N. Schmitt, "Uses and abuses of coefficient alpha," in *Psychological Assessment, 8*, pp. 350-353, 1996.

[12] X. Fan, and B. Thompson, "Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial," in *Educational and Psychological Measurement*, *61*, pp. 517-531, 2001.

[13] D. W. Zimmerman, B. D. Zumbo, and C. Lalonde, "Coefficient alpha as an estimate of test reliability under violation of two assumptions," in *Educational and Psychological Measurement*, 53, pp. 33-49, 1993.

[14] G. J. Cizek, S. Rosenberg, and H. Koons, "Sources of validity evidence for educational and psychological tests," in *Educational and Psychological Measurement*, *68*, pp. 397-412, 2008.

[15] G. J. Cizek, D. Bowen, and K. Church, "Sources of validity evidence for educational and psychological tests: A follow-up study," in *Educational and Psychological Measurement*, *70*, pp. 732-743, 2010.

[16] G. J. Cizek, "Defining and distinguishing validity: Interpretations of score meaning and justifications of test use," in *Psychological Methods*, *17*, pp. 31-43, 2012.

[17] W. Schaufeli, A. Bakker, and M. Salanova, "The measurement of work engagement with a short questionnaire," in *Educational and Psychological Measurement*, *66* (4): pp. 701-716, 2006.

[18] L. Wilkinson, and APA Task Force on Statistical Inference, "Statistical methods in psychology journals," in *American Psychologist*, *54*, pp. 594-604, 1999.

[19] B. Thompson, "If statistical significance tests are broken/misused, what practices should supplement or replace them?" in *Theory & Psychology*, vol. *9*, pp. 165-181, 1999.

[20] G. Cumming, and S. Finch, "A primer on the understanding, use and calculation of confidence intervals based on central and noncentral distributions," in *Educational and Psychological Measurement*, vol. *61*, pp. 530-572, 2001.

[21] R. E. Kirk, "Practical significance: A concept whose time has come," in *Educational and Psychological Measurement*, vol. *56*, pp. 746–759, 1996.

[22] R. Bakeman, "Recommended effect size statistics for repeated measures," in *Behavior Research Methods*, vol. *37*, pp. 379-384, 2005.

[23] S. Olejnik, and J. Algina, "Generalized eta and omega squared statistics: Measures of effect size for some common research designs," in *Psychological Methods*, vol. *8*, pp. 434-447, 2003.

[24] D. C. Howell, *Statistical Methods for Psychology*. Belmont, CA: Duxbury Press, 2002.

[25] J. H. Steiger, and R. T. Fouladi, "R2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation," *Behavior Research Methods, Instruments, and Computers*, vol. *24*, pp. 581-582, 1992.

[26] J. H. Steiger, and R. T. Fouladi, "Noncentrality interval estimation and the evaluation of statistical models," in L. Harlow, S. Mulaik, & J.H. Steiger, Eds., *What If There Were No Significance Tests?* New Jersey: Lawrence Erlbaum, 1997.

[27] 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," in *American Journal of Epidemiology*, vol. 156, pp. 1056-1061, 2002.

[28] J. Rogers, K. Howard, and J. Vessey, "Using significance tests to evaluate equivalence between two experimental groups," in *Psychological Bulletin*, vol. 113, pp. 553-565, 1993.

[29] C. W. Dunnett, "A multiple comparison procedure for comparing several treatments with a control," in *Journal of the American Statistical Association*, vol. *50*, pp. 1096-1121, 1955.

[30] C. W. Dunnett, "New tables for multiple comparisons with a control," *Biometrics*, vol. *20*, pp. 482-491, 1964.

[31] D. C. Howell, "Multiple comparisons with repeated measures," retrieved from the web on 27 Sep 2009.

[32] R. E. Kirk, *Experimental Design: Procedures for the Behavioral Sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole, 1995.

[33] S. E. Maxwell, "Pairwise multiple comparisons in repeated measures designs," *Journal of Educational Statistics*, vol. *5*, pp. 269-287, 1980.

[34] J. Jensen, D. Berry, and T. Kummer, “Investigating the effects of exam length on performance and cognitive fatigue,” in *PLoS ONE*, vol. *8(8)*, pp. 1-9, e70270, 2013.