

Measuring the Impact of Innovations in Bertheussen's 'Digital School Examinations'

Alan Fask¹, Fred Englander¹ & Zhaobo Wang¹

¹ Silberman College of Business, Fairleigh Dickinson University, Madison, NJ, United States

Correspondence: Fred Englander, Silberman College of Business, Fairleigh Dickinson University, Madison, NJ, United States. E-mail: englandr@fdu.edu

Received: August 6, 2014

Accepted: August 28, 2014

Online Published: September 25, 2014

doi:10.5539/ibr.v7n10p11

URL: <http://dx.doi.org/10.5539/ibr.v7n10p11>

Abstract

This paper reviews Bernt Arne Bertheussen's recent article published in this journal which offers a promising new approach to teaching undergraduate finance courses. Bertheussen (2014) presents a comprehensive approach which argues that a greater emphasis on the use of spreadsheets satisfies the students' desires to develop increased experience, familiarity and skill in working with information and communication technology. Such a practice, suggests the author, would also nurture a deeper, more enduring understanding of the underlying finance concepts covered. The current review of this innovation offers Bertheussen substantial credit for an important innovation in finance pedagogy. However, this review does draw attention to and analyzes the substantial limitations in the approaches that Bertheussen utilizes to measure the effectiveness of this innovation. Alternative approaches to gauge the effectiveness of this new teaching method which might serve to persuade more instructors of the utility of such a pedagogical advancement are discussed.

Keywords: digital cheating, effectiveness of digital exam procedure, experimental and quasi-experimental research designs, surface versus deep learning

1. Introduction

In his article, "Digital School Examinations: An Educational Note of an Innovative Practice," Bertheussen (2014) presented a creative approach to the use of computer spreadsheets in the teaching and assessment of student performance in finance courses taught over a nine year period at a Norwegian university's business school. For two mandatory assignments and the final examination, students are asked to consider a problem presented in a spreadsheet and use the spreadsheet to develop a solution to the problem. Even though all of the students in a given class would receive comparable problems, the problems faced by individual students were not identical because key numerical values or parameters in the problems were programmatically randomized so that each student was presented with individualized problems. Bertheussen (2014) explains that this pedagogical innovation was implemented in order to (a) respond to student preferences for an instructional approach which made greater use of information and communication technology, (b) accomplish a higher order (i.e., 'deeper') level of learning of the relevant finance topics and (c) accomplish a hindrance of student cheating on the assignments and final examination. Bertheussen (2014, p. 136) states in the conclusion of his paper that these three research questions "guided our work."

The present authors believe that Bertheussen deserves substantial credit for this thoughtful and inventive pedagogical approach. His paper offered several potentially significant contributions to the literature by using an innovative spreadsheet based format for an undergraduate finance course. These contributions included:

- Theory based design. The course was not just developed to communicate the principles of finance to the students, but to also meet important theoretical educational goals. These goals included the embedding of both formative and summative components into the course. Furthermore, the important goal of fostering deep learning was explicitly addressed.
- Use of cheating deterrents. A number of procedures, including a scheme for numerical randomization of key parameters in problem sets, were integrated into the testing process for the explicit purpose of limiting cheating opportunities.
- Novel automated test scoring. Not only was test scoring automated, but the automated scheme actually

detected consecutive errors and modified test scores appropriately.

These above contributions are substantially meritorious. Bertheussen (2014) then proceeded to offer evidence supporting the empirical validation of the methodology. It is in this evidence where the paper requires further scrutiny, for the evidence is, at best, weak. While few real life investigations rarely meet the perfect standards of experimental design as, for example, outlined in Campbell & Stanley (1963), a reasonable attempt to emulate those standards should be considered in any empirical investigation claiming validity and reliability of conclusions. It is the intention of these comments to highlight the need for validation procedures which meet reasonable standards of study design when making claims of course effectiveness and hindrance of student cheating. The Bertheussen (2014) study is thus being used as an example of common study design failures. The objections to Bertheussen's (2014) efforts to offer validation of his approach focus on several issues.

2. Fostering Deep Learning

2.1 Research Design

Bertheussen (2014) utilizes a normal distribution of grade scores as a benchmark against which one may measure the success of his innovative approach. He argues that for a testing procedure to be successful, it should not deviate from a national average distribution of grades for an undergraduate finance course. This is a startling claim. Since this course design was implemented to provide an improved vehicle for learning, what is the justification for using what presumably are inferior national teaching approaches as a target for accomplishment? Indeed, a truly effective teaching approach would result in substantial negative skewness, with the bulk of the distribution being at higher grades. For this reason, the use of the normal distribution is also suspect. Additionally, the measures of success should have been explicitly detailed prior to the beginning of the study.

A more appropriate research design would have been to randomly assign students at the outset of the implementation of this new pedagogical approach to either a treatment group or a traditional course (representing a control or comparison group). Then, at the end of the course, the investigator would compare measures of effectiveness, such as mean course grades. It is to be hypothesized that the treatment (innovative) group will demonstrate higher mean grades and that these higher grades will be statistically significant. It is rarely wise to use a national standard, whether the national standard is normally distributed or otherwise, since the subjects in the national distribution are not necessarily equivalent to those in the group under study, in this case the students in the innovative group. Treatment and control/comparison groups randomly selected from the same population would be expected to be equivalent and therefore would yield meaningful measures for comparison. Among the classic sources that outline experimental and quasi-experimental research designs to evaluate the impact of interventions such as the use of spreadsheets with individualized questions that Bertheussen (2014) considers are those by Campbell and Stanley (1963), Cook and Campbell (1979) and Shadish, Cook and Campbell (2002). In recent issues in this journal, such experimental or quasi-experimental research designs have been utilized by Mitteleman, Macieira and Avila (2014) and Seetisarn and Chiaravutthi (2011).

In establishing a course design which puts great emphasis on spreadsheet construction, Bertheussen introduces an interesting tradeoff. With more time spent on spreadsheet construction, less time may be available for the presentation of traditional financial topics. Thus the "deep learning" which Bertheussen hopes to impart with the more challenging use of spreadsheets may be offset by the lack of deep learning in the areas of traditional financial topics to which the students are not exposed. That is, the opportunity cost aspect of Bertheussen approach is not explored. Thus test measures used to gauge course learning should include both spreadsheet and more traditional questions. It would not be surprising to see that Bertheussen's students perform better on spreadsheet problems than the more traditional problems. In any case, a test-control design, which randomly assigns one class to receive the traditional course and another receive Bertheussen's course might allow for measurement of the tradeoff. The class exam, which should be equivalent for both groups, should utilize a carefully crafted set of questions, avoiding a selection of questions which were more similar to the teaching methods experienced by either group.

2.2 Measuring Deep Learning

Bertheussen (2014) claims that the fact that some students could develop spreadsheets from scratch demonstrates that deep learning was achieved by these students. Certainly, in the current business environment, where spreadsheet facility is highly desired, if not required, of new finance employees, the ability to develop spreadsheets from scratch is a much desired skill. But does the demonstration by some students of this ability truly show that the innovative course led to deep learning? That is, it seems very reasonable that there is a continuum representing the "depth" of student learning and that questions which simply rely on the rote skills of students to memorize or apply a "cookbook" procedure may lie close to "surface learning" end of the continuum.

However, Bertheussen (2014) provides no truly persuasive evidence that the individualized spreadsheet base intervention that is discussed actually represents “deep” learning. That is, he does not apply any established instrument to demonstrate that his assignment and exam questions are more appropriately directed to imposing on students a need to accomplish deep learning. He only uses reports that the students did less well on those questions which applied the intervention. While the present authors are not concluding that deep learning might not be a plausible explanation for these results, it is also possible that the pattern of findings presented could (not necessarily ‘did’) result from other characteristics of the different types of questions—e.g., differences in the clarity of the questions or differences in the time spent on the material that are linked to the questions. Bertheussen may wish to define deep learning as the ability to create a spreadsheet from scratch; however, it is a leap to believe that this ability truly measures the deep learning construct.

On a much more practical level, Bertheussen (2014) did not measure student’s skill with spreadsheet at the beginning of the course. It is quite possible that those who were able to develop a financial spreadsheet from scratch, entered the course with better spreadsheet skills and that was the principle reason for their success, rather than the innovative course or the presumed accomplishment of deep learning.

3. Hindering Cheating on Digital Exams

Bertheussen (2014) provided a number of procedures including numerical randomization in their test problems and then claims that these procedures lead to a lower level of cheating. The basis of this claim is the lack of identical test response patterns among any of the students. However, even if there had been a perfect match, this need not demonstrate cheating. If two students studied together and misunderstood some solving procedures, they might well have responded in the same incorrect way, even in the absence of cheating. Therefore, it may be argued that a similarity in the pattern of responses among students is therefore, at best, a poor indicator of cheating. A growing literature on empirical methods to analyzing cheating has been developing over the last ten to fifteen years. The studies evaluating whether online examinations are more susceptible to cheating that Bertheussen (2014) cites (Grijalva, Nowell, & Kerkvliet, 2006; Lanier, 2006; Stuber-McEwen, Wiseley, & Hoggatt, 2009), which do not offer a clear consensus, are based on anonymous surveys of students. These surveys ask the student subjects about the extent to which they may have cheated in traditional, proctored exams relative to online exams. However, it seems reasonable to question how reliable these surveys may be. Should we expect the student subjects who may have cheated to admit it, even anonymously? Student personality may play a role not only in the propensity to cheat but in the pattern of responses to such survey questions. For example, a guilt ridden student may admit to cheating behavior even in the case of a very mild infraction. Alternatively, a sociopathic student may deny any cheating behavior even in the event of severe past cheating behavior.

A second methodological approach is to attempt an experiment or a quasi-experiment involving a comparison of exam performance between two groups of classes or students—a group taking online exams and a second group taking comparable or equivalent exams in a traditional, proctored setting. This quasi-experimental approach has been utilized less often (the present authors are only aware of the research efforts of Peng (2007), Harmon & Lambrinos (2008) and Yates & Beaudrie (2009) and would offer more credible results, especially to the extent that such a study is careful to insure that the composition of the two groups and the exams that they take are equivalent. A third approach, normally offering a greater level of credibility would be based on a randomized experiment. Such a research design studying online cheating was performed by Hollister and Berenson (2009). They studied two sections taking the same business computer skills course in the same semester. They were careful to assure that the students in each section had comparable abilities and characteristics. Once the semester was underway, Hollister and Berenson (2009) randomly assigned one of the sections to a traditional assessment mode (in-class, proctored exams) and the other section to an assessment mode based on online examinations. Another such experiment has been recently undertaken by Fask, Englander and Wang (2014). Two sections of introductory statistics were assigned to complete a practice final exam several days prior to the actual final exam. A short time before these exams, one section was assigned on a random basis to take the exams in a traditional, face-to-face environment and the second class was assigned to take those exams in an unproctored, online environment. The research design allowed separate estimates to be made of the effect of the testing environment (online or in-class) on exam performance and the effect of the online environment on students’ propensity to cheat.

The above research on measuring cheating raises the question of the implementation of those approaches to Bertheussen’s method of randomization of key question parameters to avert cheating. While this approach has intuitive appeal, confirmation of its effectiveness again requires a test-control study with one group receiving randomized questions and another receiving nonrandomized questions. The same test grading algorithm

(whether automated or not) should be applied to both groups. It should be pointed out that Bertheussen is not clear as to the extent of proctoring during the exams. This is, of course, crucial information to understand cheating related issues. On one hand, Bertheussen (2014) discusses some of the previous research related to student cheating in general *and* in the case of online examinations. On the other hand, the author refers to exams in which the students “brought their own devices” (Bertheussen, 2014, p. 129) *and* states, “The final exam is not open-book ...” (Bertheussen, 2014, p. 132).

4. Other Measures of Program Success

First, Bertheussen (2014) notes that the course has existed for nine years and has been revised each year and was therefore considered a success. The longevity of the course only demonstrates the longevity of the course, not its success. Second, Bertheussen (2014) gave students a 1-5 Likert scale for two questions on their attitude towards the spreadsheet based course. The responses indicated a high level of satisfaction. This may very well be a reasonable conclusion. It should be pointed out, however, that the use of the t-distribution comparing responses (on a one to five scale) to the indifferent response (i.e., three on that scale) is problematic because the t-distribution is really not designed to be used for ordinal scales; a nonparametric approach would be preferred. It should also be pointed out that these results may well be subject to the “person-positivity bias” that pertains to such evaluations that was reported by Sears (1983). Sears (1983) found that when over 300,000 individual student research subjects at UCLA over fourteen school terms were asked to evaluate their professors on a nine point scale (where a 5 rating was labeled as average) professors were rated at 7.22 and courses were rated at 6.85. This suggests a strong positive bias in the evaluation of professors and courses which may well have applied to student satisfaction results that Bertheussen (2014) presented. Comparisons of the results to similar student feedback for similar business courses which did not involve the particular use of individualized spreadsheet assignments and exams (either in place before the intervention or for business courses at the same university) would offer stronger evidence of the positive student response to the intervention.

5. Conclusion

The above criticisms of Bertheussen’s work are not intended to question the potential pedagogical value of the innovation that he advances. This innovation appears thoughtful and promising. These criticisms are meant to highlight the point that if claims of an intervention’s success are to be made, they should be based on appropriate outcome measures that have been developed in a less ambiguous manner. The credibility of the rather impressive innovative course design which Bertheussen reported, is undercut, not enhanced, by ill-conceived measures of effectiveness. That is, Bertheussen’s ability to persuade a greater proportion of fellow finance professors to adopt a similar pedagogical approach would be substantially greater if his research design and corresponding outcome measures were more carefully constructed.

References

- Bertheussen, B. (2014). Digital school examinations: An educational note on an innovative practice. *International Business Research*, 7(6), 129–139. <http://dx.doi.org/10.5539/ibr.v7n6p129>
- Campbell, D., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.
- Cook, T., & Campbell, D. (1979). *Quasi-Experimentation: Design & analysis issues for field settings*. Boston: Houghton-Mifflin.
- Entwistle, N., & Ramsden P. (1983). *Understanding pupil learning*. London: Croom Helm.
- Fask, A., Englander, F., & Wang, Z. (2014). Do online exams facilitate cheating? An experiment designed to separate possible cheating from the effect of the online test taking environment. *Journal of Academic Ethics*, 12(2), 101–112. <http://dx.doi.org/10.1007/s10805-014-9207-1>
- Grijalva, T., Nowell, C., & Kerkvliet, J. (2006). Academic honesty and online courses. *College Student Journal*, 40(1), 180–185.
- Harmon, O., & Lambrinos, J. (2008). Are online exams an invitation to cheat? *Journal of Economic Education*, 39(2), 116–125. <http://dx.doi.org/10.3200/JECE.39.2.116-125>
- Hollister, K., & Berenson, M. (2009). Proctored versus unproctored online exams: Studying the impact of exam environment on student performance. *Decision Sciences Journal of Innovative Education*, 7(1), 271–294. <http://dx.doi.org/10.1111/j.1540-4609.2008.00220.x>
- Lanier, M. (2006). Academic integrity and distance learning. *Journal of Criminal Justice Education*, 17(2), 244–261. <http://dx.doi.org/10.1080/10511250600866166>

- Mittelman, M., Macieira, F., & Avila, M. (2014). It doesn't hurt to ask: Assessing default effects in a Brazil gas station. *International Business Research*, 7(6), 61–66. <http://dx.doi.org/10.5539/ibr.v7n6p67>
- Peng, Z. (2007). Giving online quizzes in corporate finance and investments for a better use of seat time. *The Journal of Educators Online*, 4(2), 1–18.
- Sears, D. (1983). The person-positivity bias. *Journal of Personality and Social Psychology*, 44(2), 233–250. <http://dx.doi.org/10.1037/0022-3514.44.2.233>
- Seetisarn, P., & Chiaravutthi, Y. (2011). Thai Consumers willingness to pay for food products with geographical indications. *International Business Research*, 4(3), 161–170. <http://dx.doi.org/10.5539/ibr.v4n3p161>
- Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Stamford, CT: Cengage Learning.
- Stuber-McEwen, D., Wiseley, P., & Hoggatt, S. (2009). Point, click, and cheat: Frequency and type of academic dishonesty in the virtual classroom. *Online Journal of Distance Learning Administration*, 12(3), 1–10. Retrieved from <http://www.westga.edu/~distance/ojdla/fall123/stuber123.html>
- Yates, R., & Beaudrie, B. (2009). The impact of online assessment on grades in community college distance education mathematics courses. *The American Journal of Distance Education*, 23(2), 62–70. <http://dx.doi.org/10.1080/08923640902850601>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).