UDC: 378+371.26+371.212

Dr. Nitza Davidovitch and Dr. Moshe Einat

ELIMINATING INSTRUCTOR BIAS IN GRADING: A COMPARISON OF MANUAL AND AUTOMATIC GRADING

The Ariel University Center operates an automated grading system for multiple-choice exams, which are used in many university departments in Israel. This exam method is well-known all over the world, but its innovation at Ariel University Center is the software that is used to optimize test quality and improve the professional standards of writing tests and grading students. This study examines the performance of the grading software when compared to results of manual grading. Specifically we sought to examine whether any differences would emerge in grades awarded by the software and by manual grading by instructors who also take into account the reasoning underlying students' final answers. We examined this question on a test case of an exam in the Introduction to Electrical Engineering Course at the Ariel University Center. We conclude that automated grading generates results that are closely tied to reality.

Keywords: *grading, instructor bias, multiple-choice exams, manual grading, automatic grading.*

Introduction. Technological changes, by their very nature, are designed to serve human beings and satisfy their needs; occasionally technological changes transform society and individuals. Technological inventions are assimilated into and become an integral part of the new social order. It is possible to study the assimilation of technology through research observations that focus on the effects of the new technology, compared to previous technologies. For example, with the invention of the television in the 1950s and 1960s, its efficiency as a teaching medium was examined in comparison to traditional teaching methods. Similarly, in the 1970s and 1980s, a broad range of computer-aided teaching methods was examined in a similar manner, as were multimedia applications in the 1980s and 1990s. Since the late 1990s, online learning and distant teaching have been studied comparatively, with the aim of examining their relative efficiency and effectiveness (Bernard et al., 2004).

The use of technology in academic institutions all over the world has increased significantly in recent years (Jones & O'Shea, 2004). Much effort is invested in the development of digital online learning settings, as technology is considered to offer flexibility in time, space, and pace of learning and teaching (Inglis et al., 2002). Furthermore, a series of advantages are identified with technology, including a significant improvement in the utilization of learning time, reduced learner's dependence on the site of learning (Hiltz, 1995), extension of the learning setting and information sources, elimination of dependence on textbooks as the single source of knowledge, development of an active learning environment (Hiltz, 1995), enhanced learning dialogue, economies of time and resources, and other benefits.

One of the most prevalent applications of technology in education in general, and in higher education in particular, is the use of computer software to conduct digital versions of multiple-choice tests. This test method is well known and widely used worldwide (Gamliel, 2005). Software program development produces increasingly sophisticated applications as time passes. Today, the purpose of such programs is to maximize the quality of the testing process and enhance the professional standards of the grading process. Such technological developments, and others, have created a revolution and pose a challenge for the education system in general, and higher education in particular (Leung & Ivy, 2003). These new tools require a re-thinking of the methodologies we use in academic teaching (Passig, 2003). Indeed, transferring responsibility for grading from the human factor to the technological factor offers a series of advantages: savings in time, reduced human errors, improved items based on previous experience, and other advantages. Despite the efficiency of the technological element in this process, however, the question arises as to whether, and to what degree, the testing method affects test grades.

Objective, multiple-choice tests. Multiple-choice exams are a means of assessment comprising closed items with a constant number of possible answers. Typically, such tests are used in broad examinations of existing knowledge. Multiple-choice exam scores constitute an objective means of evaluating students' mastery of the study material, and a common index for students' "extent of knowledge" of the study material.

The basic assumption is that the test items constitute a random sample of the study material, over which students are assumed to have attained mastery. Like other measurement tools of this kind (questionnaires), tests should be reliable and valid (Notzer, 2003). Test reliability is determined by calculating the weighted average of the correlations between scores calculated separately for various test sections. The test score constitutes an indirect measure of the nature of a student's knowledge, but may be influenced by many variables that are independent of the questionnaire: nervousness, guessing, physiological or health-related factors, examinee's age, personality, tester's mood, the use of different criteria determine text grades by different testers (Even Zohar, 2004), among others.

Furthermore, some items may not constitute a representative sample of the study material. For example, in the event that the items are not randomly taken from the entire body of material, but rather are concentrated in the beginning of the material, even students who do not study all the material can succeed, while students who study all the material may forget the material at the beginning and fail. If the items are taken from the end of the material, students with a short memory span can succeed even if they did not study all the material for the test.

Therefore it is clear that test scores do not have any independent existence, and do not truly reflect students' "extent of knowledge." The purpose of a test is to distinguish between those who are better or worse, and the range of test scores is unimportant. Therefore, it is more important to know each student's relative position in the group (percentile score) or her position relative to the class average (Z-score), than the student's raw test score.

From numerous aspects, multiple-choice tests are superior to the alternatives, because they support a larger sampling of knowledge items, they are more objective, and relatively easier to grade (using computer software). Open-ended questions are subject to the tester's subjective interpretations, and even in the natural sciences, mathematical solutions may be evaluated differently by different instructors (who do or do not count the method, who do or do not award points for incomplete solutions, etc.).

Scores on handwritten tests are strongly influenced by the legibility of the handwriting, the organization and neatness of the exam, and other similar factors. Multiple-choice tests offer a great degree of objectivity and gives equal opportunities to all examinees. Such tests conveniently allow instructors to mix the order of the test items to prevent cheating. Computer-aided grading saves time and prevents unnecessary arguments with students since the computer is completely "objective." Computer-aided grading makes it easy to identify test items that are overly difficult or easy, based on the number of students who succeed or fail each item, and such items can be discounted from the final test score. Furthermore, test item reliability can be easily tested (using item analysis or factor analysis), test items that do not belong to the general domain can be identified, and other actions taken to improve test scores during the grading process (Gamliel, 2005).

Performing a run while calculating test scores. Since academic institutions typically use a uniform scale that defines a passing score in the range from 60 to 100, instructors must transform the raw scores obtained on any test to a range that is close to this, using a uniform, objective formula. The score transformation must maintain the order of scores, but not necessarily in a linear manner. There are various methods for score transformation (Even Zohar, 2004). For example, a reliability test can be used with optional elimination of items that do not belong to the test domain and reduce reliability (Alpha if Deleted). In ordinary tests, however, this is not a critical issues, as the distortion caused by such items affects all examinees equally.

Instructors can examine whether a test contains items belonging to a single domain or to several domains, using factor analysis. For example, in a mathematics exam, some items may be testing for knowledge in algebra, others test knowledge in geometry, while yet others test knowledge in statistics. Instructors can calculate and award each student a separate score for each domain, and attribute different weights to the domains when calculating the final score.

It seems that in any case, it is better for instructors to give more difficult exams than easier exams. It is easier to explain a bonus score to students than to explain why his grade was reduced. The explanation that the original (raw) score has not meaning whatsoever, and that what counts is the student's position in the group, is not resistant to emotional claims of discrimination. An overly high average is just as bad as an extremely low average because the ceiling (or floor) effect—the typical upper limit of 100 or the bottom limit—prevent a normal distribution of scores with a reasonable standard deviation. A normal distribution and reasonable standard deviation are essential for any test that purports to truly distinguish between students.

Transformation should not be performed when there are a small number of examinees in the group or when the test is a make-up test (Moed bet). However, if the make-up test is very similar in format and level of difficulty to the original exam, instructors may rely on a sample of students who sat on the original exam date in determining the scores for the make-up test, by combining the scores on both tests and determining the final score on the basis of each student's position in the combined group.

Introducing a Computer-Aided Multiple-Choice Exam Grading System.

A. We conducted a workshop for teachers to teach them about developing and analyzing multiple-choice exams. Training focused on the principles of test development and utilizing the statistical properties of scores to analyze achievements. Instructors also learned how to read the software computer printout.

B. Teachers developed exams based on the guidelines covered in the workshop.

C. Students and exam administrators were briefed on the new system.

D. At the conclusion of each exam, materials were transferred to the computer center, and returned to each instructor and the exam department after grading. System experts also added suggestions on how to improve the scores. Results provided instructors with immediate feedback on the quality of their teaching, the structure of the course, and students' mastery of course assignments. The computerized analysis was accompanied by statistical data on the test's potential use as a measure of students' achievements. Instructors review the results, consult with the system expert, and may request a second run after making minor changes (such as changing an answer, adding an answer, eliminating an erroneous answer).

E. An additional run was performed for several exams. Instructors received a detailed report of the quality of the exam and suggestions for improvement in future exams.

Research questions:

1. Does the testing method affect grades?
2. What is the extent of the testing method's influence on grades?

3. How many grades will be identical when both methods are used?

4. How many grades will be different when both methods are used, and what method generates higher grades?

5. How many grades will be significantly different when both methods are used?

Method. One hundred and twenty six students sat for the exam for Introduction to Electrical Engineering, a course taught at the Faculty of Engineering, Ariel University Center. The exam comprised four questions, each question was divided into five or six sections, for a total of 21 test items. Students were explicitly and emphatically requested to provide complete solutions to the questions in the exam notebook and to also select one out of seven possible answers, corresponding to the solution they wrote. The test sheet explicitly noted that an answer without reasoning might be disqualified. Exams were graded using two separate, independent methods: computerized grading based on the answers marked on the answer sheet, and manual grading in which examiners checked each question separately and decided whether it was correct or incorrect. In most cases, this method was implemented in this manner, with the exception of single border-line cases in which partial marks were awarded for partial answers. Grades were compared and are presented below.

Findings. For the 126 examinees, the following distribution of grades was obtained for the two grading methods (see Table 1):

40% of the students were awarded the identical grade by both methods.

16% of the students completed one more answer on the computerized answer sheet compared to the number of answers they completed in the exam book.

32% of the students completed three more answers on the computerized answer sheet compared to the number of answers they completed in the exam book (of the total 21 test items).

For a small percentage of students, their computerized answer sheet contained up to 13 more answers than their exam book contained. In other words, these were students who hardly wrote anything in their exam book yet marked correct answers on the computerized marking sheet, without having any support in their exam books. A total of 18% of the students had between 4 and 13 more answers on the computerized answer sheet compared to the exam book.

The opposite pattern was also obtained at a smaller scale. In other words, there were students who completed a reasoned answer in their exam book but failed to mark the correct answer on the computerized answer sheet. Sometimes this happened due to a slight computational error, confusion, or other reasons. 9.5% of the students fell into this category. These were students who earned a higher grade in the manual grading method.

Another level on which the data can be examined is the weight of the surplus answers: in other words, the number of surplus answers multiplied by the number of students who have the same number of surplus answers. A more uniform distribution of the weight of surplus answers is evident. In other words, while the contribution of a single surplus answer is distributed over 20 students and causes a relatively small difference in the grades awarded by both methods, the distribution of 10 surplus answers, with the same weight, is distributed over 2 students only. In other words, it causes a dramatic difference in the grades awarded by both methods.

Another figure describing the entire class is the number of surplus answers per class. It is evident that the number of surplus answers per class was 233. After deducting the answers with a negative surplus, we obtain a similar figure – 215. If we attribute this number to the total possible number of answers per class (2646 answers, which is 126 students × 21 questions), we see that less than 9% are surplus answers on the computerized answer sheet. Their distribution is presented in the above table and figure.

Table 1

Distribution of exam results

Number of extra answers on the computerized answer sheet compared to the exam notebook	Student number	%	Weight	Total
0	51	40.5%		
1	20	15.9%	20	
2	8	6.3%	16	
3	12	9.5%	36	
4	3	2.4%	12	
5	4	3.2%	20	
6	4	3.2%	24	
7	4	3.2%	28	
8	4	3.2%	32	
9	0	0.0%	0	
10	2	1.6%	20	
11	0	0.0%	0	
12	1	0.8%	12	
13	1	0.8%	13	233
-1	7	5.6%	-7	
-2	4	3.2%	-8	
-3	1	0.8%	-3	-18
	126	100.0%	215	215

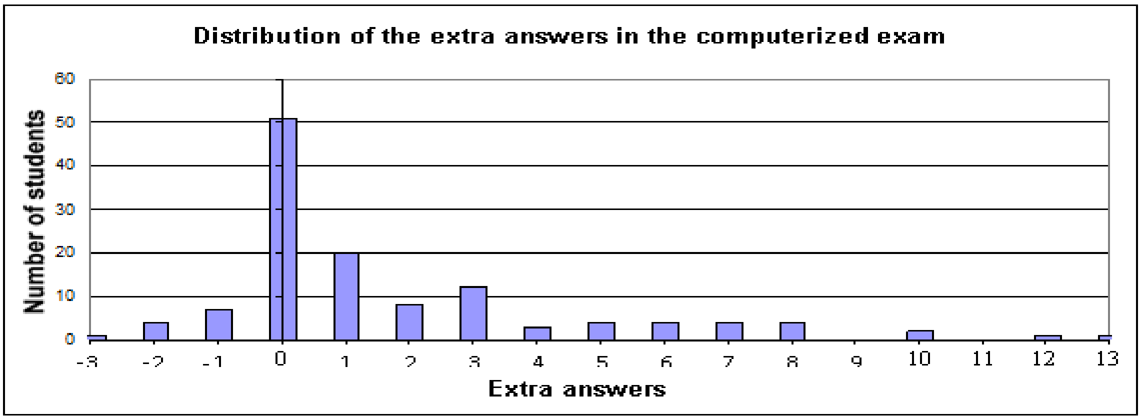

Figure 1. Distribution of exam results. Students' grades are presented by the number of extra answers on the computerized answer sheet (compared to the number of answers in the exam notebook). Approximately 40% of the students received the same grade using both methods.

Following are two additional data that offer perspective on the findings:

A. Mixing – The computerized exam sheet was mixed using 'Test Perfecto', a software program that automatically mixed the items on each exam sheet. This prevents students from blindly copying answers from other students, because no two exam sheets are identical. Students may still try to obtain the correct answers from other students, but the opportunity for cheating in this manner is limited. Therefore, although the surplus answers may be attributed to cheating, but only in a limited manner.

B. Despite the instructions that were given explicitly to students, some students complained after the exam that they calculated their answers on their calculator or on the edge of the page and did not bother to copy the entire solution into the exam notebook. Therefore, some of the surplus answers may be attributed to this phenomenon.

Summary and discussion. In the present study we sought to examine the advantage of technology over man, and visa versa, in academic teaching practice, on a test case of exam grading. Specifically we suggested that there would be differences in grades when exams were graded manually and automatically. Student findings indicate that technology, despite its many advantages, is unable to maintain the same level of authenticity and responsibility as instructors who manually check and grade exams, using their discretion, addressing students' reasoning and process, and evaluating the picture in entirety.

Findings demonstrate that both methods generate somewhat different grades – grades were higher using the automated grading system compared to manually graded exams. For 71% of the students, the difference was less than 10%, for 16% of the students, the difference in grade was 20% or more, in favor of the automated system. Total surplus answers in the automated method were only 9% but this is not distributed uniformly. Clearly, some students did not bother to enter part of their answers into their exam notebooks, and this explains the findings.

The idea for this study was promoted by this phenomenon precisely – very high grades in a subject that is known for its high level of difficulty. Study findings show that computerized grading tends to award a higher grade in 48% of the cases. This teaches us about the potential implications of a shift to technology. Traditional test grading was within the instructor's sphere of responsibility but has become a technological task – instructors submitted their students' answer sheets to the Exam Unit, which enters them to the computer, which generates the examinees' grades. Eliminating the human factor from the equation has risks, in addition to its advantages.

In contrast to the human instructor, the computer does not apply any discretion. The only thing it generates is correct/incorrect results. Checking exams based on "the bottom line" accounts for the objectivity of the grading process, but this is a slippery slope: The absence of human intervention may lead to loss of control over one of the most important processes in teaching – assessing students' achievements. We believe that the instructor, the person who is responsible for students' learning process, should be involved in the end result. We teach our students values that emphasize, reasoning, effort, and process, yet automated grading ignored such considerations and limits students' abilities to a single number in a box.

Nonetheless, automated grading generates results that are closely related to reality, assuming that reasonable efforts are made to prevent cheating which is, as well-known, easier in computerized exams. Although the instructor loses the personal contact to the exam results, and has the fear regarding the reliability of the computerized exam, the difference in the results in this case study is comforting. There is a difference, but it is limited to less then 10%. Automated grading tends to favor students relative to manual grading that involves discretion and possibly subjectivity. This tendency toward a higher grade, in our opinion, points to future research directions. It is the tendency of the Western world to rely totally on computers and their performance. The results of this study indicate that despite the precision of the automated grading method, this is nonetheless precision of the end result only. As teachers and educators, part of the "higher education system" we wish to teach our students that the end result, although important, is not everything. Intellectual efforts, creativity, in-depth understanding, intent – all these are worthy of recognition and appreciation. These values, in our opinion, cross the boundaries of specific subjects and with time become values for life. In the rapidly changing post-modern world that demands results here and now, we must emphasize to our students the importance and value of the way. The way, even if it does not lead to the correct answer, is worthy of recognition. The opposite is also true – students who fail to present the reasoning or an in-depth explanation for their answer should not be given the positive feedback embodied in a high grade. All this on behalf of those values that cross over the boundaries of the exam and continue to affect ethical values that are worthy of being taught to the generation of the future.

REFERENCES

1. Bernard, R., Abrami, P., Lou, Y., Borokhovski, E., Wade, A., & Wozney, L.,. (2004,). How Does Distance Education Compare With Classroom Instruction? A Meta-Analysis of the Empirical Literature. *Review of Educational Research*, *74*, 379-439 [in English].

2. Even Zohar, S. (2004). Transformation of exam grades. *Meida Bareshet*, Computer Center Newsletter, Bar Ilan University, Issue 10. Retrieved from: http://www.biu.ac.il/Computing/meyda/dec_2004/tranziun.shtml [in Hebrew].

3. Gamliel, A. (2005). On the meaning of grades in institutions of higher education. *Al Hagova, 4*, 42–43 [in Hebrew].

4. Hiltz, S., R., (1995). *Teahcing in a virtual classroom*. International Conference on Computer Assisted Instrucion ICCAI 95 [in English].

5. Inglis, A., Ling, P., & Joosten, V., (2002). *Delivering digitally*. London: Kogan Page [in English].

6. Jones, N., & O'Shea, J., (2004). Challenging hierarchies: The Impact of e-learning .*Higher Education* 48: 379–95 [in English].

7. Leung, Y., L., & Ivy, M., I., (2003). How useful are course websites? A study of students' perceptions. *Journal of Hospitality, Leisure, Sport & Tourism Education, 2*, 15-24 [in English].

8. Notzer, N. (2003). Writing multiple-choice exams. *Al Hagova*, *2*, 42–43 [in Hebrew].

9. Passig, D. (2003). A taxonomy of future higher thinking skills. *Informatics in Education – An International Journal, 2,* 79–92 [in English].

Ницца Давидович, Моше Эйнат

УСТРАНЕНИЕ ПОГРЕШНОСТЕЙ ПРИ ОЦЕНИВАНИИ: СРАВНЕНИЕ ТРАДИЦИОННОГО И АВТОМАТИЗИРОВАННОГО ОЦЕНИВАНИЯ ЗНАНИЙ СТУДЕНТОВ

Ариельский университетский центр применяет систему автоматизированного оценивания для единого государственного экзамена, который проводится во многих университетах Израиля. Этот экзаменационный метод известен во всем мире, но Ариельский университетский центр применяет инновационную программу, которая позволяет оптимизировать качество оценивания и проверки знаний студентов. В данной статье изложены результаты исследования качества работы данной программы в сравнении с традиционным оцениванием знаний студентов. В частности, авторы статьи исследовали разницу между конечными оценками работ студентов, произведенными программой и человеком. Исследование, выполненное на материале результатов экзамена по предмету «Введение в электрическую инженерию» в Ариельском университетском центре, позволило сделать вывод о том, что результаты автоматизированного оценивания близки к объективным.

Ключевые слова: оценивание, неточности оценивания, единый государственный экзамен, традиционное оценивание, автоматизированное оценивание.

Ніцца Давидович, Моше Ейнат

УСУНЕННЯ ПОХИБОК ПІД ЧАС ОЦІНЮВАННІЯ: ПОРІВНЯННЯ ТРАДИЦІЙНОГО І АВТОМАТИЗОВАНОГО ОЦІНЮВАННЯ ЗНАНЬ СТУДЕНТІВ

Аріельскій університетський центр застосовує систему автоматизованого оцінювання для єдиного державного іспиту, який проводиться в багатьох університетах Ізраїлю. Цей екзаменаційний метод відомий у всьому світі, але Аріельскій університетський центр застосовує інноваційну програму, яка дозволяє оптимізувати якість оцінювання та перевірки знань студентів. У статті викладені результати дослідження якості роботи цієї програми в порівнянні з традиційним оцінюванням знань студентів. Зокрема, автори статті досліджували різницю між кінцевими оцінками робіт студентів, виробленими програмою і людиною. Дослідження, виконане на матеріалі результатів іспиту з предмету «Введення в електричну інженерію» в Аріельскому університетському центрі, дозволило зробити висновок про те, що результати автоматизованого оцінювання близькі до об'єктивних.

Ключові слова: оцінювання, неточності оцінювання, єдиний державний іспит, традиційне оцінювання, автоматизоване оцінювання.

Подано до редакції 08.09.2014