

Randomised Items in Computer-based Tests: Russian Roulette in Assessment?

Anthony M. Marks

Faculty of Engineering, Information Technology & the Built Environment, Summerstrand North Campus
Nelson Mandela Metropolitan University // anthony.marks@nmmu.ac.za

Johannes C. Cronje

Faculty of Informatics and Design, Cape Peninsula University of Technology // johannes.cronje@gmail.com

ABSTRACT

Computer-based assessments are becoming more commonplace, perhaps as a necessity for faculty to cope with large class sizes. These tests often occur in large computer testing venues in which test security may be compromised. In an attempt to limit the likelihood of cheating in such venues, randomised presentation of items is automatically programmed into testing software, such that neighbouring screens present different items to the test-taker. This article argues that randomisation of test items can be a disadvantage to students who were randomly presented with difficult items first. Such disadvantage would violate the American Psychological Association's published guidelines concerning testing and assessment that call for the principle of fairness for test-takers across diverse test modes. Owing to the smallness of the chance of a student being randomly assigned difficult items first, it may be hard to prove such disadvantage. However, even if only one test-taker is affected once during a high-stakes test, the principle of fairness is compromised. This article reports on four instances out of about 400 in which students may either have been unfairly advantaged or disadvantaged by being given a series of easy or difficult items at the beginning of the test. Although the results are not statistically significant, we conclude that more research needs to be done before one can ignore what we have named the Item Randomisation Effect.

Keywords

Computer-based tests, Fairness, Cheating, Randomisation of items, Anxiety

Introduction

An important security feature of computer-based multiple-choice testing is that test items are randomised to prevent students working at adjacent computers from copying. The downside of such randomisation, however, is that it prevents planned sequencing of items, which is commonplace in the paper-based equivalent. A test constructor may, for instance, place easier items at the beginning of the test to build confidence in the test-taker, and place the most difficult items at the end so that slower students' time is not wasted by attempting items that are beyond their ability. In other cases, a student may choose to go through the test first and select the easier items, leaving the difficult ones for last. Randomising items does not accommodate a test user or a constructor who wishes to ensure that items progressively become tougher. Although navigation through computer-based tests is possible, it is certainly an inconvenience to test-takers who want to leave tougher items for last. In this study we wanted to know if it would be possible to identify students who were disadvantaged in a test because they had been randomly assigned the difficult items first.

The purpose of this study is to determine the effect of current computer-based testing practice on student performance, particularly with respect to the random item sequencing algorithm. The purpose of this algorithm or software code is to ensure that each test-taker is administered the test items in a sequence different from that of any other test-taker. This effect will not affect all test-takers equally and is like a form of Russian roulette, a dangerous game of chance, played in high-stakes testing contexts. Russian roulette does not cause injury to all the participants, yet is deadly to the one who pulls the trigger last. One test-taker affected by this form of unfairness in assessment is one too many. Test security may be enhanced, but is test fairness for some test-takers compromised?

Paper-and-pencil tests still account for a major portion of any student's final result. This is true from the early school years to the final years of postgraduate studies, during which time one can safely estimate that at least ten years of school achievement was assessed with the aid of written tests and examinations. Good test-taking habits will ensure that candidates perform according to their level of preparedness (Glenn, 2004). The test-taking skill of particular interest and relevance to this paper is the one that requires candidates to select their perceived easy questions first

and answer these before attempting their perceived tougher questions (“Use Parent Nights,” 2004). In paper-based tests, this is easily achieved by simply paging back and forth through the test.

Now with the advances in personal computer technology and huge investments in evaluation and testing software (Billings, 2004; Harding, 2001; Varughese, 2005), computer-based testing is becoming commonplace. Candidates who have good test-taking skills with regard to paper-based tests will still outperform candidates without these skills in computer-based tests. Or will they? Algorithms that randomise the order in which the test items are presented to each candidate automatically control certain computer-based test assessments.

Although randomisation may reduce the security risk (Pain & Le Heron, 2003) of adjacent students’ copying from or aiding one another, it may unfairly increase test anxiety for some of the candidates. Navigation through numerous test items in a computer-based test is not conveniently achieved by candidates, so randomly receiving several difficult items consecutively may unduly stress a candidate. Increased anxiety at any stage during the test for whatever reason is likely to have a negative effect on that person’s performance for the remainder of the test (Lufi, Okasha, & Cohen, 2004; Supon, 2004). Obviously, the sooner the sequence of difficult items is presented the more pronounced the effect it may have upon the remainder of the test. The main question of this study is “Can instances be found where randomisation of items in a computer-based test unfairly disadvantaged any of the test-takers in any way?”

The following sub-questions drove the study:

- What constitutes “normal performance” for each student in the sample?
- Did any candidates’ performance differ significantly from their “normal performance”?
- Were these candidates presented with consecutive randomly sequenced difficult items?
- Can this deviation be attributed to the randomisation of items presented to the student?

For the purpose of this study, the authors felt that primacy and recency effects on memory and cognition were not directly relevant. Primacy is a term from cognitive psychology that is used to explain the increased likelihood for accurate recall of the initial items of a list of items. Recency is a term used to explain the increased likelihood for people to accurately recall the items occurring at the end of the list. Studies of these effects (Bemelmans, Wolters, Zwinderman, ten Berge, & Goekoop, 2002; Talmi & Goshen-Gottstein, 2006) are more concerned with understanding the workings of memory and recall and offset discussion around long-term and short-term memory processes. However, in this study, a set of test items comprise different multiple choice questions, not a list of words and or numbers to be recalled in the correct sequence. The items or questions in this study are randomised so that test-takers are administered the same test, but the randomised sequence of items may subtly cause some candidates to be disadvantaged in comparison to the rest of the candidates taking the test.

The focus of this study is on the potential threat to fairness that randomisation of test items may cause to any one test-taker. This is especially important as item randomisation is done in high-stakes testing scenarios, for example, entrance examinations to tertiary institutions, promotion or retention of children in schools, or selection of potential candidates for a particular job vacancy (Russell, Goldberg, & O’Connor, 2003; Zenisky & Sireci, 2002). Consider what it would be like to be the test-taker and have the items presented in a manner that causes you to perform poorly in such an assessment, and to be disqualified from further consideration for the institution/post for the wrong reason. Hence, even if only one test-taker is unfairly administered a test, it is one too many.

Literature survey

Tobias (1985) hypothesized that test anxiety reduced the cognitive ability required to solve problems, thus leading to poor results. On the other hand, a student with good test-taking skills needed less cognitive capacity to spend on the physical elements of the test and could therefore concentrate more on recall of the actual learning content. Bierbaum (2007) built upon Tobias’s work and identified a perceived alignment between instruction and assessment. Bierbaum argued that students come to expect a certain style of test and answer in a certain way. Deviations from such expectation lead to test anxiety.

This study seeks to show that a form of test anxiety may exist in computer-based tests due to randomisation of the items. This randomisation may affect a small percentage of test-takers in a sample and therefore may be seen as unfairly disadvantaging those candidates.

Fairness for everyone taking tests is the underpinning principle of the various publications on testing and assessment available from the American Psychological Association (APA, 2004; Turner, DeMers, Fox, & Reed, 2001). These guides are specifically used to ensure standardised tests and interpretations of test-taker abilities made from such tests are accurate and fair (Turner et al., 2001). Test developers and users are defined in the guidelines, and they are regarded as the stakeholders specifically tasked with ensuring that the guidelines are followed, and that fairness for all test-takers is achieved. Another publication has evolved from the guides to inform all three stakeholders, viz. developers, users, and test-takers, of the rights and responsibilities of test-takers (APA, 1988).

From the literature, fairness is seen to be a fundamental principle. Randomised sequencing prevents cheating, and test-taking skills involve a number of activities, some of which are compromised by randomised sequencing. Test stress negatively impacts test results, while navigational control enhances test performance. No literature could be found that comments specifically upon the sequencing of computer-based test items. However, Sternberg (1998) stresses the importance of metacognition as a part of what makes an expert student. We could argue that randomised sequencing impairs metacognition because it distracts from the holistic nature of a test.

Pain and Le Heron (2003) reported that randomised sequencing of test items was successful in preventing cheating in computer-based tests. In one of the scenarios, they even allowed the computer-administered test to randomly select different items from the question databank such that each student had a different collection of items presented to them for the assessment. Not surprisingly, this ensured that no students could cheat, but as the authors reported, the concern about the fairness of such a solution is questionable and they returned to allowing random sequencing of a set of test questions, such that all test-takers essentially took the same test. However, Pain and Le Heron did not consider the harmful effects of unfairly presenting sequences of difficult items early in the test.

Glenn (2004, p. 62) advises test takers to “answer the easiest questions first. Completing the sure-thing questions first boosts student confidence from the outset.” The literature does not say what happens to student confidence or anxiety if this important test-taking skill is ignored. We would like to deduce that the opposite effect will occur, that student confidence wanes and test anxiety increases. In computer-based testing contexts, several consecutive difficult items presented to a test-taker will increase that test-taker’s anxiety level. Therefore, that candidate is unfairly administered (Turner et al., 2001) the test in comparison with all the other candidates that were fortunate not to be randomly presented with such a sequence of difficult test items. This is clearly in contravention of the APA requirements that all tests are to be administered in a standardised manner such that all test-takers are given an equal opportunity to provide evidence of their abilities in that test (2004).

The above assumes that a sequence of difficult items will indeed cause increased anxiety. It also assumes that increased anxiety has a negative effect on student performance for the remainder of the test (Black, 2005). Cognitive ability decreases during states of tension and increased anxiety (“Reduce Test Anxiety,” 2005; Dutke & Stöber, 2001; Hancock, 2001). A review of literature pertaining to computer-based tests found various studies relevant to computer anxiety (Lufi et al., 2004; Supon, 2004; Tseng, Tiplady, Macleod, & Wright, 1998) and how it adversely affected student performance. Computer anxiety is prevalent in persons who seldom use computers in their daily lives, hence their nervousness when using computers. This computer anxiety compounds the natural anxiety caused by the need to perform adequately in a test setting, as stated by Bugbee Jr. (1996, Specific Research Studies section, para. 17): “Anxiety is quite real and can gravely affect a test taker. It must be dealt with.”

Various studies have already managed to identify diverse factors that cause significant differences in student performance to be observed across modalities, specifically paper-based versus computer-based test modes (Bugbee Jr., 1996; Carlson & Smith Harvey, 2004; Hoff, 1999). These differences are solved directly. For example, the ability to return to any item and edit its answer has helped ensure some equivalence across test modes (Ferguson, Kreiter, Peterson, Rowat, & Elliott, 2002). If a solution is not easily implemented, then the APA guidelines allow for scores to be adjusted such that the adjusted scores are fair representations of the test-takers when compared with persons taking tests in other test modes (Russell et al., 2003). However, the item randomisation test mode effect identified by this study has not yet been studied.

The paper and pencil mode conveniently allows test-takers to read through the entire set of items before choosing to attempt the easier ones first, as prescribed by those advocating this approach as a good test-taking tactic (“Use Parent Nights,” 2004; Glenn, 2004; Staber & Pekrun, 2004). However, randomised sequencing of test items in computer-based test assessments is not determined by the test-taker. Navigation between items is not as convenient for computer-based test assessments as it is for paper and pencil, and this forms the basis of the argument that students in computer-based tests are all at a relative disadvantage. This navigation mode effect has been studied (Ferguson et al., 2002), and as all test-takers in this modality are equally affected, it was readily shown that by allowing candidates more control, the disadvantage could be somewhat negated. However, it then became obvious that students then needed extra time to navigate back and forth through the test items as compared to students sitting the same test in a paper-based mode.

The mode effect described in our study will only affect a small number of test-takers randomly in a computer-based test, not all of the candidates. It follows that this effect is not likely to be easily noticed. Hence the need for this study, as the available literature does not include any studies in this regard.

Method

Overview

In this study we investigated the performance of 103 third-year students of veterinary science in four tests out of five that were presented during a year-long course. These tests were completed in the year prior to the research commencement as it was hoped that individual occurrences of the randomisation effect would be found from existing data. From an ethical perspective it is important to note that no students were disadvantaged as a direct result of this investigation. This research is an ex-post facto study done on existing computer-based test data. Further, the research is a pilot study done to investigate the potential or need for rigorous experimental research studies to be designed to properly investigate what has now been called the Item Randomisation Effect. No students were interviewed before, during, or after the collection of the data as this was outside of the intended scope of the pilot study. Similarly none of the tests was manipulated in any way. The data was simply collected from the computer-based testing department at least one year after the tests were administered as this was believed to be sufficient for the purposes of this pilot study. The data were analysed and the normal performance of each student was established as described below. The test data were processed until the relevant variables were ready for analysis. Next, the students who deviated significantly from their normal performance were flagged, and their test experience was then analysed in detail. The graphical representation of selected students is included later in the paper. It must be stressed that this research was ideographic rather than nomothetic — we were looking for specific instances rather than considering the over-all test performances.

Normal performance for this study

The performance of a candidate who was presented with difficult items during the initial stages of any of the tests needs to be compared with the normal performance of that candidate in similar assessments, and significant deviations in the performance of that candidate need to be reported. One obvious record of normal performance could be obtained from each student’s academic record and the average obtained thus far in his or her academic career. However, this was not satisfactory because deviations above and below the average are normal, based upon the individual student’s aptitude and motivation levels for different courses. We decided, rather, to find a sample of students who had sat more than two computer-based test assessments in one year-long course. The average obtained for each candidate for the particular course would, for the purpose of this study, be considered the normal performance of the student.

Defining difficult items for this study

The Difficulty Index and the Discrimination Index are two of the indices that are readily available in computer-based tests. Most testing software calculate these as standard features to assist users in designing and retaining quality test items, yet flag problematic items that should either be edited or discarded (Alessi & Trollip, 2001; Reise & Henson,

2003). In the context of this study, the difficulty index is the obvious data that allows one to check the correlation between the sequence of items and the relative difficulty of those items. The discrimination index is useful but of lower significance for the purposes of this pilot study, and is therefore not considered any further in this paper. The difficulty index is calculated by dividing the number of correct responses for an item by the total number of attempts made to answer the item (Reise & Henson, 2003). The index can range between zero and unity. A zero index value means that none of the test-takers could correctly answer the item. A difficulty index of unity means that all the test-takers attempting the item chose the correct option (or key). Assuming the items were not compromised in any way, a zero index implies a difficult item and, conversely, a unity index implies an easy item. Once used in a test, each item can now be ranked according to its difficulty index and thus labelled and stored in the question databank.

Sample data selection for this study

The computer-based testing section of the University of Pretoria assisted us in finding a set of data that satisfied the requirements of numerous tests in one subject in a particular year for as large a sample of students as possible. The chosen sample consisted of four sets of test results for 103 veterinary students in their third year of study in 2004. One of their modules required them to sit four computer-based test assessments. Questionmark 3.2 was the software used to administer the tests. The data available for each test consisted of four text files generated by Questionmark after the completion of each test. In addition, we were supplied a portable document format (pdf) copy of each set of test items. Permission to use the data was granted by the Veterinary Sciences Faculty. Considerable effort was required to organise the four text files such that the data required for this study could be analysed. This is because the required variables appeared in different text files.

Data cleaning

Four sets of existing test data from a sample of students for one course in their academic year was used for this study. They are labelled Test 1, Test 2, Test 3, and Test 5. Test 4 data was eliminated because problems during the test led to its being postponed until the next day. The test items were all changed as many of the candidates had seen some of the items, so a new test, Test 5, was created. Thus only four sets of test data were useful.

The normal performance for each test-taker was taken as the average scored for the four tests in this course, as was explained earlier. Next, it was important to obtain the sequence of items as randomly presented to each test-taker. The last step required the difficulty index of each item to be recorded adjacent to each item for the candidate. The student score for each item was also included as was the discrimination index and other potentially useful information. The process was repeated for each of the four sets of test data. The data was then ready for analysis.

Discussion from findings

Detailed analysis of the particular students who had test scores varying by at least 15 percent does seem to show correlations between the sequences of item difficulty and the effect on students' performance on the test. The literature indicates that beginning with easy items in a test is one habit likely to have a positive effect on student grades (Glenn, 2004). In this study, students who scored significantly higher than usual in Test 1 tended to be presented with the easy items early in the test. Conversely, those presented the difficult items early in the four tests tended to perform significantly below their normal levels.

The figures that follow, one from each test, are the most striking examples of the trends that become apparent when one studies the data pertaining to the candidates who showed a significant difference in performance from their normal performance for this module across the four tests. The following conventions are followed in the figures:

- (1) Diamonds indicate the difficulty index for each item
- (2) The horizontal line labelled "average difficulty" is also the class average scored for the particular test
- (3) Diamonds 10% above the line of average difficulty are considered "easy"
- (4) Diamonds 10% below the line of average difficulty are considered "difficult"
- (5) Diamonds within 10% of the line of average difficulty are considered "average"
- (6) The candidate's deviation from normal is indicated under the chart's title.

Figure 1 shows an example of a candidate who scored 24 percent above his/her normal. Closer inspection of the chart shows that for the first eight items, six were easy, one difficult, and one of average difficulty. This candidate was presented the items in an almost ideal sequence, easy to difficult, and scored significantly above his/her normal, which supports the literature pertaining to attempting easy items first (Glenn, 2004; Supon, 2004).

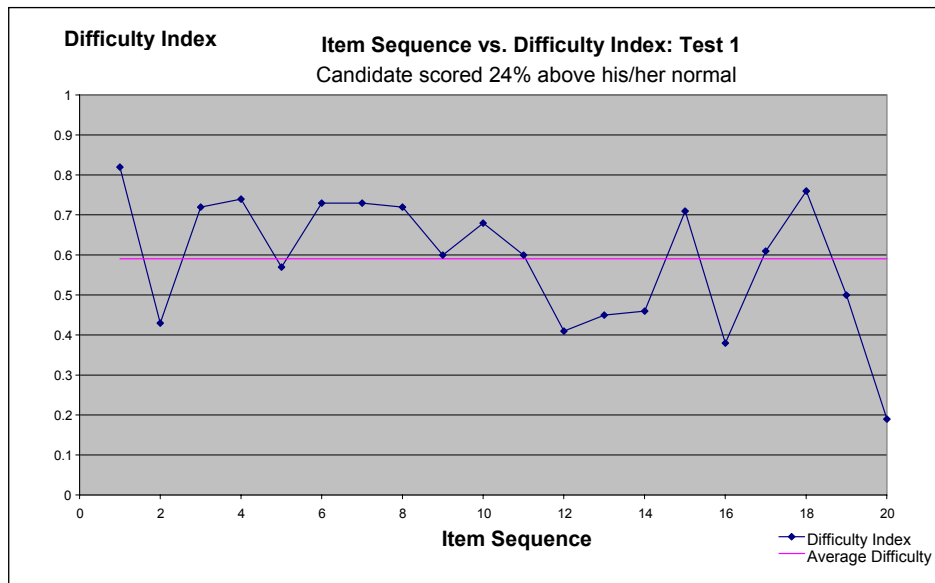


Figure 1. Example from Test 1

Figure 2 shows an example of a candidate who scored 19 percent below his/her normal level. For the first eight items, two were easy, two difficult, and four of average difficulty. The ideal would have been easy items in the beginning, average items in the middle, and difficult items at the end. It is important to remember that the assumption that the difficulty index indicates the degree of difficulty is valid for the class group as a whole but not necessarily true for each individual student. This particular student got the second and fifth items correct, inferring that the other six items of the first eight were perceived to be difficult for this particular student, hence the potential for increased anxiety as described throughout this paper.

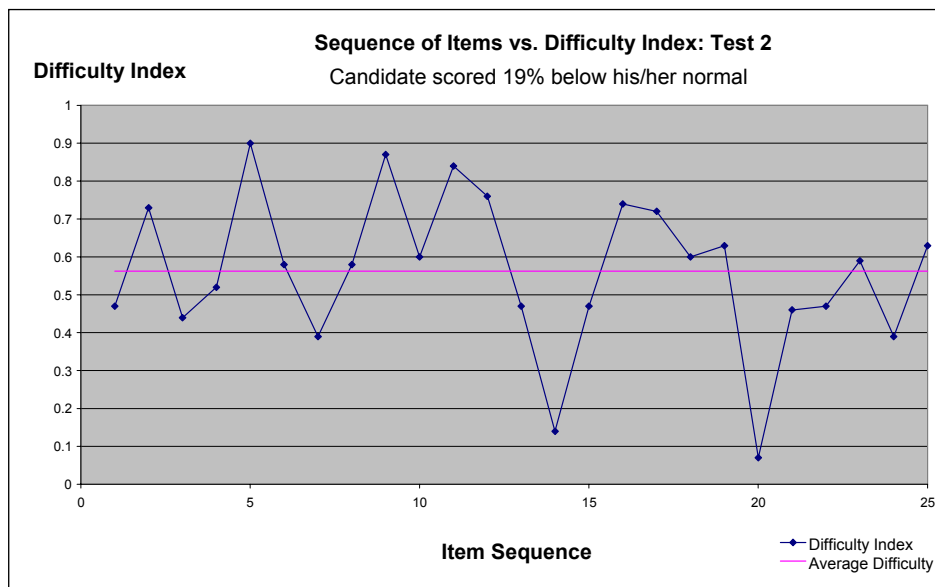


Figure 2. Example from Test 2

In Figure 3 the candidate scored 15 percent below his/her normal level. In first third of the test (the first 16 items), four questions were easy, six difficult, and six of average difficulty. Most of these questions ranged from average difficulty to difficult, which is ideal for the middle third of the test. Did this random sequence of test items cause increased anxiety?

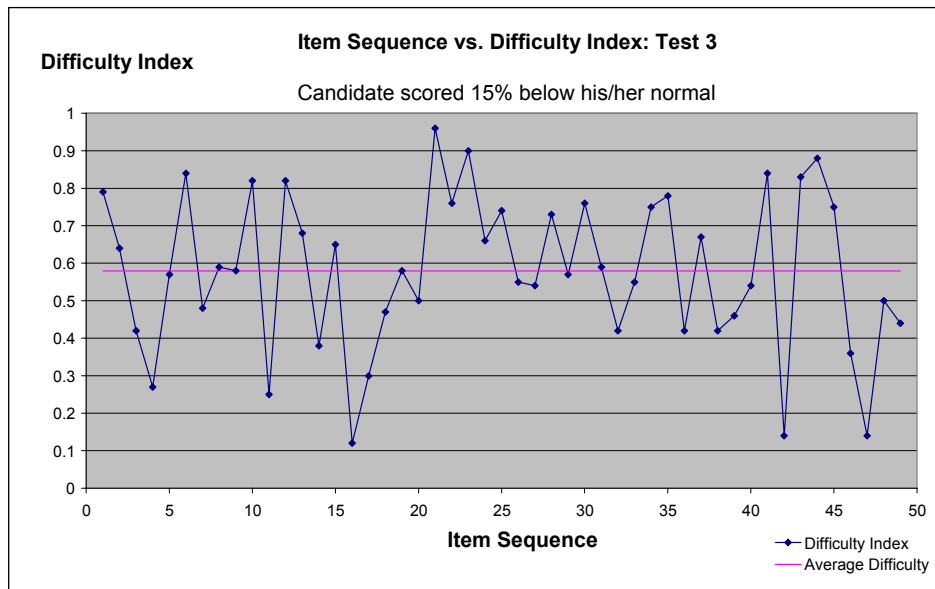


Figure 3. Example from Test 3

Figure 4 shows a candidate from the fourth test who scored 15 percent below his/her normal. Of the first sixteen items, three were easy, four difficult, and nine of average difficulty. Again, this shows that the initial third of the test was of average difficulty, which is ideal for the middle portion of the test, but not for the initial portion of the test. Even closer inspection shows only six items above the average line and ten items on the difficult side below the average line. The initial items presented to this candidate clearly tended to be the more difficult items. This indicates to us that this student too could have been a victim of unfair assessment by having been randomly presented with difficult items early in this test.

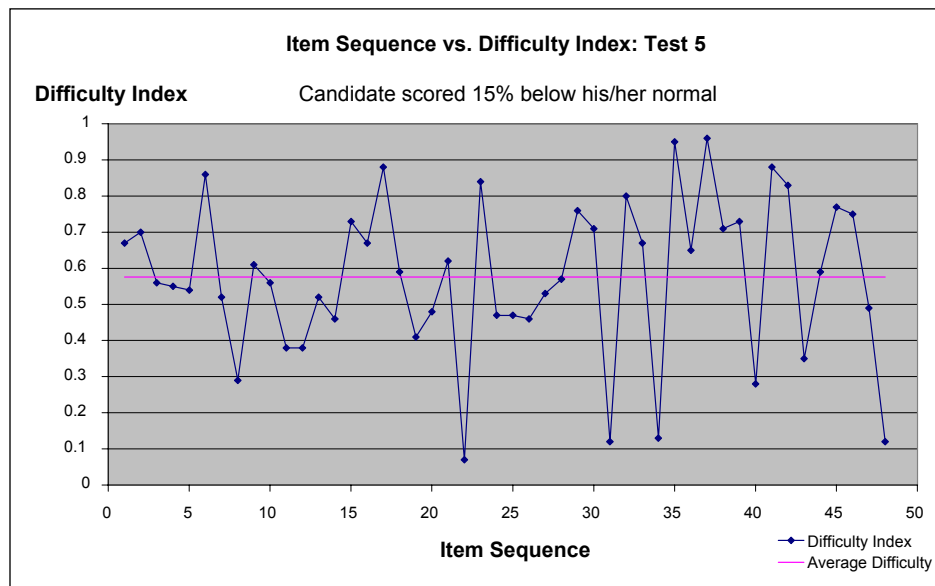


Figure 4. Example from Test 4

Limitations

The sample is far too small for such a study, and therefore one can only conclude that an experimental study must be specifically designed, in which all the candidates in the experimental group are presented several difficult items consecutively, and all candidates in the control group are presented with several easy items consecutively, with the results of the two groups then compared.

Formal statistical analysis was conducted on the cleaned data, more in an effort to cover all the bases and not because any significant correlations were expected to be found for this ideographic investigation. As was expected, no statistically significant differences were found. However, the infrequency of the affected test-takers in the sample does cause the statistical software to generate a warning that “chi-square may not be a valid test.” Table 1 below illustrates the problem with analysing the data. For each of the tests, fewer than ten percent of the sample scores vary by at least 15 percent from the test-takers’ normal performance, calculated as explained earlier in this paper. In fact, the percentage drops below five for the last two tests. This is mainly due to the decision to use existing computer-based test data; however, a properly designed experimental research project would ensure the study becomes nomothetic, and hence statistical analysis becomes useful for the investigation of potential correlations.

Table 1. Scores varying by at least 15% from normal

	Test 1	Test 2	Test 3	Test 5
Scoring lower	2	5	3	2
Scoring higher	7	2	0	0
Total	9	7	3	2
% of Sample	8.7%	6.7%	2.9%	1.9%

The above limitations are explained further in terms of construct validity in Bugbee Jr. (1996), with regard to the equivalence of tests across modes. An error of measurement (error variance) is contrasted with systematic variance with respect to the mode of administration. The important distinction is in noting that systemic variance affects all test-takers equally, whereas “error variance or error of measurement is variation of errors due to chance” (Bugbee Jr., 1996, Specific Research Studies section, para. 12). This study is typical of error variance, and error variance in this context shows that either the variance is not due to the mode of administration, or that presenting difficult items early will likely cause as much increased anxiety in paper-based tests too. We are not contradicting this, but are concerned that the effect is more pronounced in computer-based test assessments due in part to the lesser convenience in this mode for navigating through the test items. One way around this dilemma is to design a test across modes that ensures that all candidates are presented with the difficult items early such that the assessment gets progressively easier. The research design ensures to some extent a systemic variance in that all test-takers may be affected equally. Care must be taken to account for those candidates with good test-taking skills and low levels of test anxiety who will probably attempt the items in a progressively more difficult sequence, influencing the reliability of the study.

Conclusions and recommendations

Despite the limitations of this study, the findings are noteworthy because a potentially unfair testing practice has been identified and verbalised. Also, a potential gap in the literature can now be filled as researchers investigate the test-mode effect that, in this study, is referred to as the Item Randomisation Effect.

Unfairness in assessment is not an acceptable practice for test developers, users, or takers. The randomly affected test-taker is the one who suffers any consequences of this practice, yet it is within the powers of the test developers to ensure this won’t happen by programming algorithms in computer-based testing software. In addition, it is the responsibility of test users to ensure developers are made aware of this test mode effect. One test-taker affected by this unintended Russian roulette in assessment is one too many, as it would violate the APA guidelines on fair testing (APA, 2004).

Assuming this mode effect is found to cause some students to be disadvantaged in computer-based test assessments, we recommend that software vendors add a few lines of code to the randomising algorithm, such that this test mode effect is automatically prevented from occurring in future computer-based test assessments. This is easily achieved

once items have been used in a test, as each item can now be ranked according to its difficulty index, labeled, and stored in the item database. This could be useful for ensuring that randomising algorithms present items to students randomly while progressively increasing the difficulty of the items. Randomisation to ensure test security, yet progressively allowing items to become more difficult as the test items are presented to each test-taker, will prevent occurrence of the item randomisation effect.

It is hoped that this paper will inspire researchers to create studies specifically suited to obtain correlations that will confirm or refute the item randomisation effect introduced in this paper.

Acknowledgements

Thank you to the University of Pretoria, Computer Testing Section, for their assistance and for supplying the test data; the Faculty of Veterinary Sciences for permission to use their data; and the Department of Statistics, particularly Mrs. Jaqui Sommerville, for the formal data analysis.

References

- American Psychological Association. (1988). Rights and responsibilities of test takers: Guidelines and expectations. Retrieved January 16, 2006, from www.apa.org/science/ttrr.html.
- American Psychological Association. (2004). Code of fair testing practices in education. Retrieved January 16, 2006, from www.apa.org/science/fairtestcode.html
- Alessi, S. M., & Trollip, S. R. (2001). *Multimedia for learning: Methods and development* (3rd ed.). Boston: Allyn & Bacon.
- Bemelmans, K. J., Wolters, G., Zwinderman, K., ten Berge, J. M. F., & Goekoop, J. G. (2002). Evidence for two processes underlying the serial position curve of single- and multi-trial free recall in a heterogeneous group of psychiatric patients: A confirmatory factor analytic study. *Memory, 10*(2), 151.
- Bierenbaum, M. (2007). Assessment and instruction preferences and their relationship with test anxiety and learning strategies. *Higher Education, 53*(6), 749–768.
- Billings, K. (2004). Online assessment: Perspectives of developers. *Media & Methods, 40*(4), 26.
- Black, S. (2005). Test anxiety. *American School Board Journal, 192*(6), 42.
- Bugbee Jr., A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education, 28*(3), 282.
- Carlson, J. F., & Smith Harvey, V. (2004). Using computer-related technology for assessment activities: Ethical and professional practice issues for school psychologists. *Computers in Human Behavior, 20*(5), 645.
- Dutke, S., & Stöber, J. (2001). Test anxiety, working memory, and cognitive performance: Supportive effects of sequential demands. *Cognition & Emotion, 15*(3), 381.
- Ferguson, K. J., Kreiter, C. D., Peterson, M. W., Rowat, J. A., & Elliott, S. T. (2002). Is that your final answer? Relationship of changed answers to overall performance on a computer-based medical school course examination. *Teaching & Learning in Medicine, 14*(1), 20–24.
- Glenn, R. E. (2004). Teach kids test-taking tactics. *Education Digest, 70*(2), 61–63.
- Hancock, D. R. (2001). Effects of test anxiety and evaluative threat on students' achievement and motivation. *Journal of Educational Research, 94*(5), 284.
- Harding, R. (2001). What have examinations got to do with computers in education? *Journal of Computer Assisted Learning, 17*(3), 322.
- Hoff, D. J. (1999). Testing. *Education Week, 19*(2), 6.

- Lufi, D., Okasha, S., & Cohen, A. (2004). Test anxiety and its effect on the personality of students with learning disabilities. *Learning Disability Quarterly, 27*(3), 176.
- Pain, D. E., & Le Heron, J. L. (2003). WebCT and online assessment: The best thing since SOAP? *International Forum on Educational Technology & Society, 6*(2), 62–71.
- Reduce test anxiety to improve student performance. (2005). *Teaching Professor, 19*(9), 5.
- Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*(2), 93.
- Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: A look back into the future. *Assessment in Education: Principles, Policy & Practice, 10*(3), 279.
- Staber, J., & Pekrun, R. (2004). Advances in Test Anxiety Research. *Anxiety, Stress & Coping*, p. 205.
- Sternberg, R. J. (1998). Metacognition, abilities, and developing expertise: What makes an expert student? *Instructional Science, 26*(1–2), 127–140.
- Supon, V. (2004). Implementing Strategies to Assist Test-Anxious Students. *Journal of Instructional Psychology, 31*(4), 292–295.
- Talmi, D., & Goshen-Gottstein, Y. (2006). The long-term recency effect in recognition memory. *Memory, 14*(4), 424.
- Tobias, S. (1985). Test anxiety: Interference, defective skills, and cognitive capacity. *Educational Psychologist, 20*(3), 135–142.
- Tseng, H.-M., Tiplady, B., Macleod, H. A., & Wright, P. (1998). Computer anxiety: A comparison of pen-based personal digital assistants, conventional computer and paper assessment of mood and performance. *British Journal of Psychology, 89*(4), 599.
- Turner, S. M., DeMers, S. T., Fox, H. R., & Reed, G. M. (2001). APA's guidelines for test user qualifications. *American Psychologist, 56*(12), 1099.
- Use parent nights to improve student test-taking skills. (2004). *Curriculum Review, 43*(5), 6.
- Varughese, J. A. (2005). Testing, testing. *University Business, 8*(4), 59.
- Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 326–337.