UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Random question sequencing in computer-based testing (CBT) assessments and its effect on individual student performance

by

Anthony Michael Marks

Submitted in partial fulfilment of the requirements of the degree

Magister Educationis

In

Computer- Integrated Education

Department of Curriculum Studies

Faculty of Education

University of Pretoria

August 2007

Supervisor: Prof. Dr. J.C. Cronjé

Contents

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

Summary

This research is important because it has identified a gap in the existing knowledge base. A term is therefore coined to label a computer-based test mode effect, the so-called **Item Randomisation Effect**, discussed in detail in this thesis. Item Randomisation Effect is a test mode effect occurring in computer-based testing contexts, especially noticeable in test-takers that may be susceptible to test anxiety. The practise of randomising multiple choice items in computer-based test venues is commonplace, mainly as a deterrent for cheating. Previous research attempted to determine the degree of equivalence across testing modalities of any test. The need was to ensure test-takers in paper-based tests would not have an advantage/disadvantage over test-takers given the same test in a computer-based mode. Such studies have a nomothetic perspective. This research contrasts with those earlier studies in that it has an ideographic perspective because it is concerned with the performance of individuals taking any test in the computer-based modality. This subtle difference in perspective may account for the apparent gap in the existing educational research literature. Evidence of Item Randomisation Effect was found in this study but further research into this test mode effect is necessary.

Key Words

Cognitive workload, Computer-based, Item Randomisation Effect, Primacy, Recency, Test developer, Test-taker, Test user, Test anxiety, Test-taking skills.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

1. Chapter 1: Introduction

1.1. The Study Question

The original proposal for this dissertation was presented to my supervisor and although the idea was accepted, it was too ambitious for the purposes of this qualification and so it was agreed that this should be a pilot study. The results of the pilot study would then determine if there was a need for further research to be done. Notably, this pilot study highlights a potential gap in the existing literature on assessment, and presents two arguments to illustrate this. Perhaps the advances in personal computer technology, specifically in the use of computerised assessments, may account for the apparent gap in the literature that I have identified. Although literature on test-taking skills and equivalence of testing across modalities exists, none seems to have the particular angle that I consider, and/or they have not yet been applied to computer-based testing.

Written examinations and tests, although still predominantly paper-based, 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 academic achievement was assessed with the aid of written tests and examinations. It is possible that well prepared students are getting results significantly lower than they deserve because of deficient test-taking skills. Now with the advances in personal computer technology and huge investments in evaluation and testing software, together with the advantages of immediate feedback and automatic assessment, computer-based testing is becoming commonplace. Candidates that have good test-taking skills with regards to paper-based tests are still going to outperform candidates without these life skills. Or will they?

Algorithms that randomise the order in which the test questions are presented to each candidate automatically control certain computer-based test assessments. Although this may reduce the security risk of adjacent students

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

copying from or aiding one another, it may unfortunately unfairly increase test anxiety for some of the candidates. Increased anxiety for whatever reason is likely to have a negative effect on that person's performance for the test. The two arguments presented in Chapters 2 and 3 build on this concern. Chapter 2 considers the Item Randomisation Effect (Marks, Cronje, & Mostert, in press) from a test anxiety perspective. Chapter 3 builds on Chapter 2, but focuses on possible increases in test-taker fatigue and cognitive workload caused by the Item Randomisation Effect described in Chapter 2.

This randomisation of multiple-choice question items in present computer-based testing practice is the main focus of this study. The main question that this investigation asks of society is an ethical one.

Is it morally and ethically acceptable to ignore factors causing increased test anxiety if it only affects a small number of test-takers in computer-based assessment, or is one adversely affected student one too many?

1.2. Background

I am often asked how I came to conceptualise the arguments contained in this pilot study, and I give the narratives that follow because they helped crystallise the thoughts over several years of practice as an educator of future electrical engineers.

Whilst revising for the Government Ticket of Competency (Mines & Works, Electrical & Mechanical Engineering) examination I frequently sought assistance from a colleague that was also revising for this tough examination. This person always knew the answers to my questions, hardly even needing to refer to textbooks or notes. We were both surprised that he failed, quite badly, even though I passed. He was the better-prepared candidate, and he was convinced of a successful result when he left the examination venue. At a social function weeks after the results were published, I asked him if, "Examination Technique", (at the time this is the term I used to describe "test-

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

taking skills"), meant anything to him? It didn't! Suddenly, it made sense. My colleague, who deserved to pass, had scored significantly below his true worth because of neglecting to master the skills or habits needed to write examinations well. Although adequately prepared for the content assessed, he was not equipped with the test-taking skills to ensure the evidence of learning got transferred onto the answer script, and thus failed unnecessarily.

This narrative from my past is relevant because I have since taken up a lecturing post and I notice that many of my learners seem to be unaware of the benefits of "good" test-taking skills. During invigilation of tests and examinations I notice that many candidates don't have clocks or watches. Yet they are expected to pace themselves and not spend too much time on any one question. Then during the marking process I also noticed that many learners answer the questions in numerical order, instead of beginning with their "favourite" questions first (Glenn, 2004; Supon, 2004). This attests to the probability that they are not "planning" their paper before starting. In turn this may also be because they did not read through the question paper before commencing with the answers in numerical order. "So what?" the uninformed person may ask. Do these habits really have a significant impact on the student's results?

The impact on performance and anxiety could be huge; depending on how "student friendly" the examination has been designed. If the examiner was inexperienced and ordered the toughest questions to appear first and gradually allowed the questions to get easier (a common fault of even experienced examiners), then I propose that candidates neglecting to read and plan the sequence in which they attempt examination questions, could obtain a fail grade for that paper. This actually happened to my top student when I, as an inexperienced examiner, set an examination paper that got progressively easier. The majority of my students passed that paper well, but they started with the last and easiest question (Glenn, 2004). My best student completed the examination by attempting the questions in numerical order and only obtained 30% for the examination. He got 55% below his semester mark of 85%. The examination's last question happened to be the easiest but

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

he didn't even attempt it because he got stuck on the first few questions that were also the most difficult questions. The average score by candidates that attempted the last question was 20 out of a possible 25, with several of the better students obtaining the maximum of 25 for the question. The evidence on paper gets marked, the unwritten potential to answer the question is not marked, and this contributed significantly to this top student scoring below his worth. This story adequately illustrates that good habits during written tests do matter.

The discerning reader will also note that the responsibility to have good test-taking habits is not just that of the student. Examiners should also attempt to set tests in such a manner that test anxiety is likely to be reduced not increased. In other words attempt to set a test that makes candidates appear to have good test-taking skills. This is why I am concerned that a computer algorithm randomly generates the sequencing of questions presented to candidates. If even one student is getting most of the difficult questions first, s/he may obtain a score far below his/her true worth for that test.

The responsibility of the examiner to set tests that are likely to minimise test anxiety must now also be shared by the computer-based testing software vendor. The algorithm used to randomise the multiple-choice questions presented to each candidate must ensure as far as is practicable, that test items tend to get progressively more difficult as the questions are presented. This can then make up for some students' lack of test-taking skills, helping them to relax and build confidence with each correct answer, hopefully relaxing enough that recall of learning becomes possible to answer the later more difficult items.

1.3. Statement of Purpose

The purpose of this study is to determine the effect of present computer-based testing practice on individual student performance, particularly with respect to the random item sequencing algorithm used in computer-based

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

tests. Two arguments of the effect of item randomisation in computerised testing are given in Chapters 2 and 3 of this dissertation.

1.4. Research Perspectives and Designs

1.4.1. Focus of this Research

Test-taking skills can be seen from two perspectives. The one perspective is that the teacher/educator/facilitator/examiner, (the test user), is responsible to ensure a test is prepared that allows the candidates to show what they have actually learnt. The other perspective is that the child/learner/student/candidate, (the test-taker), is responsible to ensure the test is approached in such a way as to ensure he/she provides the necessary evidence of the learning he/she has acquired. The lists that follow are my opinion of what needs to be the accepted responsibilities of the obvious stakeholders in ensuring tests are able to measure the true worth of the test-taker. The points below are supported by the literature and where relevant I will cite authors that support the points. However, this is not an exhaustive list, and there are more that could be added but for the purpose of this study the list is sufficient.

1.4.2. Educator Perspective (Responsibility)

It is the responsibility of the educator to:

- Ensure learners are properly equipped with the necessary test-taking skills (Glenn, 2004; Supon, 2004).
- Ensure tests are set in a way to maximise learner achievement by minimising learner anxiety.
- Ensure learner fatigue is minimised by test layout (paper-based) and minimised scrolling on screen (computer-based tests) (Ricketts & Wilks, 2002).
- Ensure the learner has had sufficient time to practise and master the tools required to provide the evidence of learning, whether for paper-

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

based or computer-based tests (Russell, Goldberg, & O'Connor, 2003).

- Feel relaxed (not anxious) before, during and after the test (Black, 2005; Hancock, 2001).

1.4.3. Test-taker Perspective (Responsibility)

It is the responsibility of the candidate to:

- Practise the test-taking skills and master them ("Use parent nights to improve student test-taking skills," 2004).
- Practise anxiety management techniques to ensure maximum brain functionality (Black, 2005).
- Score the maximum he/she can considering the preparation or study that he/she have done (Glenn, 2004).
- Feel confident that he/she can use the tools supplied in either a written paper-based test or the software in a computer-based test.
- Feel relaxed (not anxious) before, during and after the test (Black, 2005; Hancock, 2001).

1.5. Main Question and Sub-questions

Is it morally and ethically acceptable to ignore factors causing increased test anxiety if it only affects a small number of test-takers in computer-based assessment, or is one adversely affected student one too many?

- Does present computer-based testing practice actually affect student performance? Is there really a problem?

- Can small adjustments to present computer-based testing practices have a significant positive impact on student results in these tests?

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

- Can the affects be explained in terms of increased test anxiety or cognitive loading?

1.6. The Rationale and Purpose of this Study

Most educators expect tertiary students to master test-taking skills during the scholastic period of their academic career, and therefore assume that this skill has been acquired. My observations at some South African tertiary institutions indicate that these skills have in fact not yet been learnt by most students. Conversely, it is also reasonable to expect that an examiner with even a few years of experience at setting written assessments would learn to set fair tests/examinations that obtain valid and reliable scores representing each student's worth. In South African tertiary institutions, such expectations are far from being realised because many students actually lack the necessary test-taking skills, and lecturers are too preoccupied with research outputs to be bothered with the finer details of assessment practice. Our students are scoring approximately fifteen to twenty percent below their worth in each test and examination that they sit for due to poor test-taking skills (students' responsibility) and perhaps even lower, up to fifty percent, if this factor is compounded by poor test-setting skills (examiners' responsibility). Well-designed research needs to be undertaken in order to prove or disprove my opinion. This pilot study will hopefully be used to justify a properly designed PhD research project to investigate the above opinions.

The results of this research will be of benefit to students, educators and even to those holding the purse strings at our educational institutions. The first two role players are to benefit in obvious ways, but the financial manager is not an obvious beneficiary. Presently in South Africa, our government subsidy formulae payouts are based on students that pass or are successful. The institution receives greater monetary funding if more students pass. If my hypothesis is proven correct, students may all score one to two symbols higher without having to study harder or drop standards. All stakeholders need to become aware of this limitation and to properly see that the

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

necessary test-taking skills are learnt and used (students and examiners are jointly responsible).

The above rationale is relevant to all modalities of assessment; however, this pilot study is limited to the computer-based modality. Primarily this is because computer-based testing is a convenient vehicle to use to quantify the extent to which performance is affected due to the automatically generated psychometric data and statistics that become available to researchers after each computerised test.

1.7. Objectives of this Study

The first objective is to show that there is indeed a possibility that individual students may be affected by the randomisation of multiple-choice items in a computer-based test. This is achieved in Chapter 2 and in Chapter 3.

Secondly, I aim to link the findings of this pilot study to literature that may be used to guide future research into this particular area of assessment practice.

1.8. The Justification of this Study

Every year South African tertiary institutions turn prospective learners away due to the poor symbols attained in the final year of school particularly for the Mathematics, Science and English subjects. If one assumes that poor test-taking skills result in school leavers scoring fifteen percent below their true worth in each written test and examination, it becomes obvious that more school leavers would be allowed entrance to tertiary institutions if this deficiency is addressed (see Figure 1 overpage). This may partially alleviate the problem of insufficient numbers of students qualifying for entrance to scarce skills qualifications such as Engineering, where several would-be engineering students are disqualified due to the poor results obtained in their final examinations at school level.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

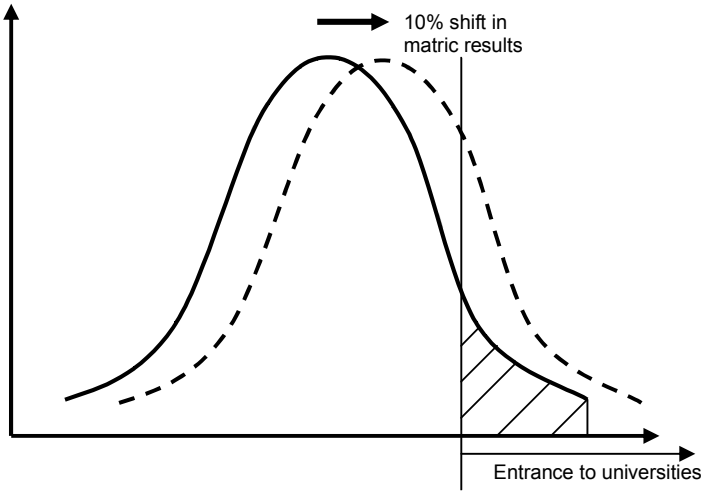

Figure 1: Increase of those gaining university entrance

The above figure (Figure 1) illustrates the possible increase in eligible candidates for entrance to university in South Africa if all learners presently practising poor test-taking skills scored higher in their school careers due to improved test-taking skills. This is my opinion based on observations of the lack of test-taking skills prevalent in the tertiary institution at which I have been employed, since 1995. The hatched area under the solid bell curve is the representation of the number before acquiring the recommended test-taking skills. The area to the right of the vertical line and under the dashed bell curve represents the greater number of potentially eligible learners if test-taking skills were improved.

The figure that follows (Figure 2) could be used to show that the quality of students eligible for qualifications with limited intakes could be improved. Consider if universities are only able to allow a fixed number of learners to enrol in any one qualification, i.e. only the best 50 students may enrol any year for a Fine Art qualification. Now from Figure 2 assume person "x" is the 50^{th} person accepted into the qualification, and person "y" is perhaps the 57^{th} on the list. Assume person "x" applies good test-taking skills in school leaver examinations/tests, but person "y" has not. If person "y" applied good test-taking skills and scored a symbol higher, s/he may have been admitted as say number 43 on the top 50 list, bumping person "x" down to person number 51,

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

who is now excluded in favour of the better quality student person “y” that was initially excluded due to poor test-taking skills.

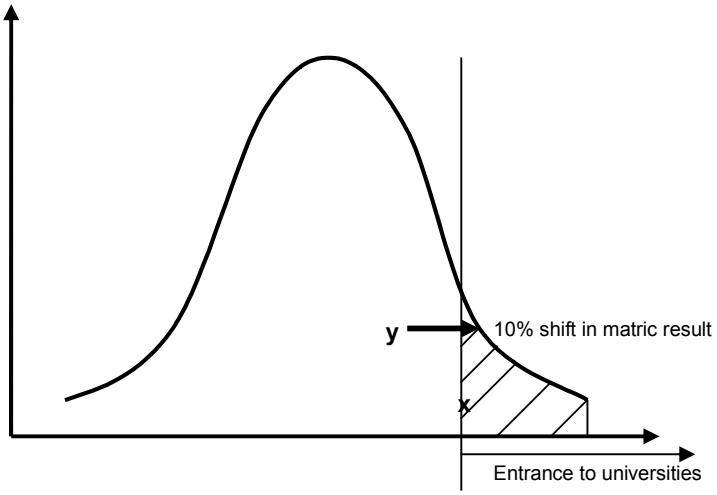

Figure 2: Improved quality for each intake

Figure 3 below can be used to illustrate the potential increase in earnings from present government subsidies for tertiary institutions. This is due to

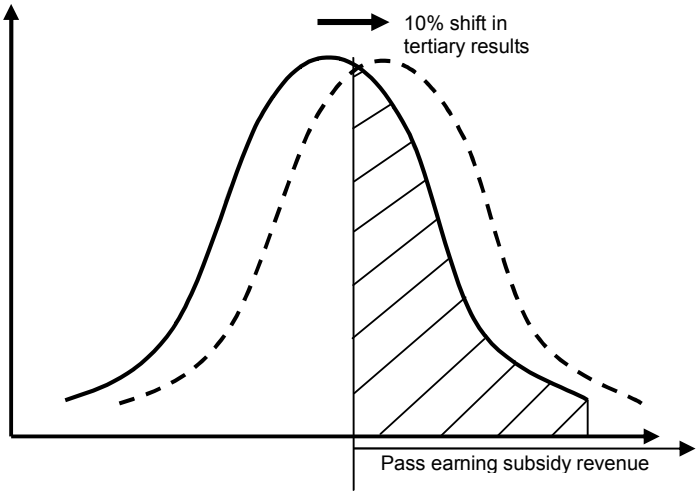

Figure 3: Increased government subsidy earnings

incentives for increased pass rates as evident in the subsidy formulae. Much pressure is placed on educators to work harder and longer to ensure that higher percentages of students pass each course without compromising present academic standards. The results of this research could alleviate some of this pressure especially if examiners set “student friendly”

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

assessments that do not compromise standards but partly compensate for students with poor test-taking skills. Assisting learners to master test-taking skills would also positively impact on pass rates without compromising standards. Many more learners would pass, if they now scored ten percentage points higher in each assessment. Revenues received from subsidy grants would increase noticeably. Students too would benefit by entering society and earning a salary earlier, as even one course repeated will result in a six to twelve month delay in qualification from the institution and the costs of repeating a course would not be unnecessarily incurred. Consider Figure 3 above for the area under the dashed curve, but to the right of the vertical line representing students passing and earning the institution government subsidy monies. Compare this to the smaller hatched area. Perhaps the improvement could be achieved by adjustments in the test-taking and /or test-setting skills of the stakeholders, the students and/or examiners at the institution concerned.

All of the above motivations are significant reasons for research to be conducted into the broader issue of test-taking skills and test-setting skills. This pilot study is relevant because the computer-based testing modality is becoming more commonplace at tertiary institutions. Yet the common practice of allowing question sequence to be randomised by a software algorithm to minimise the possibility of cheating in computer testing venues may negatively impact the test-taking skills of students and the test setting skills of lecturers. This is highlighted in Chapters 2 and 3.

1.9. Research Design

This research is a pilot study. It has been conducted on existing data collected during computer-based tests, and is therefore ex post facto analysis of existing data. Lastly it should be noted that this research is not nomothetic but ideographic, as it is concerned with the individuals taking the test, not with the performance of the test per se.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

This study used existing test data from previous computerised assessments to analyse the performance of students versus the difficulty level of questions presented to them randomly in these computerised tests. The aim was to obtain the correlation, if any, between the sequencing of the questions and individual student performance. The achievements of the students in a set of four computer-based tests were summarised into a tabular format. Those students that scored significantly higher or lower than their average for the four tests were flagged. These flagged students' tests were analysed to see if the particular test items were presented in a sequence that could account for the deviation from the norm. The sensitive and confidential nature of students' academic records meant that this information was not made available for this pilot study. Therefore it was decided to use the average of the four tests as the normal level of performance for each student for the purposes of this pilot study.

1.10. Participants/Sample and Instrumentation

Permission was granted to use the 2004 computer-based test data for 103 Veterinary students (see Appendix 1). It was important for the sample size to be as large as possible to improve the possibility of obtaining statistically significant correlations of the individuals that may have been affected by the sequence of questions presented in the tests. The sample group used for this study was the largest available at the University of Pretoria that had multiple test data results spanning an entire year of academic assessment in a single subject for the individuals in the sample. It was therefore used in this pilot study with the hope that it would prove to be large enough to obtain statistically significant correlations. Data sets for each of four computer-based tests were analysed.

In order to compare individual test scores to overall performance I compared each student's score to the average that student obtained for the four computerised assessments in that subject for 2004. The analysis and

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

interpretation of the above data is included later in the dissertation in the two arguments presented in Chapters 2 and 3.

1.11. Data Analysis

The Statistics Department assisted me in this regard. The concern for this study is that the statistical analysis may be meaningless due to the ideographic nature of this study of a nomothetic group or sample. Statistics tend to identify patterns that are relevant to larger portions of a sample, not suited to studies dealing with the exceptions of a sample. This study is by its very nature attempting to identify a problem that may not be affecting large portions of a sample. The study is concerned with the potential effect of item sequencing on any one test-taker, and whether it could be a factor in that particular person's performance in that particular test. The top student I referred to in my background, performed consistently until the particular test scenario in which he performed significantly worse than normal. Statistical analysis may not pick up this exception, as this exception would affect the standard deviation from the norm and not look for this as the rule. An error of measurement (error variance) is contrasted with systematic variance with respect to the mode of administration (Bugbee Jr, 1996, Specific Research Studies section, para. 12). The important distinction is in noting that systemic variance affects all test-takers equally, whereas according to Bugbee Jr. *Error variance or error of measurement is variation of errors due to chance.* Nevertheless the statistical analysis was done in order to be thorough though the ideographic nature of this study indicates that statistically significant correlations would likely not emerge from the analysis.

1.12. Definition of Terms

Computer adaptive testing: This refers to computer administered tests that adapt to the skill level of the test-taker (Alessi & Trollip, 2001). The test items are selected randomly from a large databank and are ranked

created using BCL easyPDF Printer Driver
Click here to purchase a license to remove this image

according to difficulty. The test items presented are designed to pinpoint the test-taker's ability in a relatively short time, often requiring less than ten items to be presented to any one test-taker. This study is not concerned with computer adaptive testing per se.

Computer anxiety: A feeling of nervousness occurs for some persons when required to use a computer. This is quite debilitating in these individuals and can prevent them from performing at their normal levels of ability.

Computer-based testing: This refers to testing contexts administered using a computer in a venue designed for this purpose, often using software specifically developed to administer tests to test-takers. This type of testing tends to make use of multiple-choice type questions, but is not limited to this type of question.

Computer test-taking anxiety: This is not necessarily experienced by persons that normally experience nervousness when using computers, but may be a nervousness experienced by individual test-takers when required to perform in a computer-based test assessment, that do not usually get anxious when administered assessments in other testing modalities such as oral or paper-and-pencil tests.

Examination: Also known as an exam (USA) or test or assessment.

Examiner: Person/s responsible for setting tests and assessments. The examiner is also one of the stakeholders referred to in this study as a *test user*, (see test user).

Ideographic perspective: Is concerned with the performance of individuals taking the test, and how their performance is impacted through the Item Randomisation Effect. This is the perspective taken in this study.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

Nomothetic perspective: Is concerned with the performance of the test as administered in the computer-based modality, to be sure it will be seen as equivalent to the same test when administered in a different test mode. Literature cited in this paper tended to have a nomothetic perspective.

Performance level or true worth: This is difficult to define, as it is influenced by many factors, such as student preparation, fatigue, etc. In this particular pilot study the true worth of any student was taken as the average score for any one student for the four tests making up the data set for the study. Perhaps performance level or true worth could be coupled to a student's overall academic performance in all subjects for future research studies.

Student friendly test: A test that presents the easy questions in the beginning and gets progressively tougher. This will compensate for students that lack good test-taking skills, especially those skills that enable the student to attempt the easiest questions first.

Test developers: *Test developers are people and organisations that construct tests, as well as those that set policies for testing programmes* (APA, 2004). Note this definition is not specific to any modality of testing but is relevant to computerised modalities just as it is to the paper-and-pencil testing modality.

Test-setting skills: Lecturers or examiners should become better skilled at setting tests so as to always minimise test anxiety and thereby maximise student performance without compromising the standard of questions presented to test-takers.

Test-taker: Any person that is undergoing an assessment, including scholars, students, and even persons applying for certain types of employment. Although this study is mainly concerned with computer-

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

based test assessments, test-taker is not necessarily limited to this modality of testing.

Test-taking skills: Good test-taking skills enable a candidate to consistently provide written evidence of the learning that has occurred. Poor test-taking skills result in a candidate consistently neglecting to provide the written evidence of the learning that has occurred. It is the written evidence that is now being assessed, so the candidate scores below his/her true worth due mainly to poor habits or a lack of test-taking skills.

Test users: *Test users are people and agencies that select tests, administer tests, commission test development services, or make decisions on the basis of test scores. Test developer and test user roles may overlap* (APA, 2004). This definition is also applicable across the testing modalities.

1.13. Study Aims Summarised

This study sought to show that the random sequencing of questions presented to students in computer-based tests might cause the performance of that student to be adversely affected. The effect may be due to either increased test anxiety as the more difficult questions are presented first to a particular student, or it may be due to the increased fatigue of the student as s/he attempts to navigate back and forth through the questions in an attempt to answer the easier questions first. This could place an extra cognitive load on the student that increases the test fatigue experienced by that student. These arguments are adequately presented in Chapters 2 and 3.

Finally, it is important to note that this study does not investigate the effects of computer-based testing on students with respect to race, colour, creed or gender. Such concerns, however important they may be to the reader are

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

outside of the intended scope of this study. It is also important to note that this research is limited to concerns regarding computer-based testing, and the study findings might not be applicable to computer adaptive testing contexts.

1.14. Literature Survey

1.14.1. Fairness

This paper is one that at its core is mainly concerned with assessment practice. This can be narrowed further to a concern with the principle of "Fairness" of assessment. The South African Qualifications Authority (SAQA) has registered several "Unit Standards", specifically dealing with assessment, for South African educationists to adhere to. The following is an extract from one of the unit standards' purpose statements (SAQA, 2007). *This unit standard is for people who design assessments to facilitate consistent, credible, reliable, fair, and unbiased assessments of learning outcomes.* The last page of this unit standard goes on to define what it refers to as "Principles of assessment", and here "Fair" is defined as follows:

> *Fair: The method of assessment does not present any barriers to achievements, which are not related to the achievement of the outcome at hand.*

The SAQA unit standard is very clear that the principle of fairness holds for the individual candidate just as it does for the entire group of candidates, and stresses the importance of ensuring even the special needs of candidates must be taken into account. Assessment Criterion 3 of the unit standard states that *Potential unfair barriers to achievement by candidates are identified and the design addresses such barriers without compromising the validity of the assessment or possibilities for continued learning.* The arguments presented in this paper are concerned with the potential for an unfair sequence of multiple-choice items to be administered to a test-taker due to the randomisation of items in computer-based testing contexts, referred to as Item Randomisation Effect.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

Internationally various professional bodies have concerned themselves with this principle of fairness of assessments. Perhaps the most often cited is the American Psychological Association (APA). The APA have published a code titled *Code of Fair Testing Practices in Education* (APA, 2004) and was prepared by the Joint Committee on Testing Practices. The first paragraph of the code is quoted here in its entirety because it is so relevant to the arguments presented later in this dissertation.

> *The Code of Fair Testing Practices in Education (**Code**) is a guide for professionals in fulfilling their obligation to provide and use tests that are fair to all test takers regardless of age, gender, disability, race, ethnicity, national origin, sexual orientation, linguistic background, or other personal characteristics. Fairness is a primary consideration in all aspects of testing. Careful standardization of tests and administration conditions helps to ensure all test takers are given a comparable opportunity to demonstrate what they know and how they can perform in the area being tested. Fairness implies that every test taker has the opportunity to prepare for the test and is informed about the general nature and content of the test, as appropriate to the purpose of the test. Fairness also extends to the accurate reporting of individual and group test results. Fairness is not an isolated concept, but must be considered in all aspects of the testing process* (APA, 2004).

The concept of fairness is the primary principle in the APA **Code**, and the **Code** is quite clear in the implication that fairness extends to each and every test-taker, and is to be used by stakeholders to *improve their testing practices*. The **Code** defines two stakeholders (other than the test-taker) and gives guidelines to each separately, even though it acknowledges that the roles of the two stakeholders may overlap. First, Test Developers are defined as *people and organizations that construct tests, as well as those that set policies for testing programs.* Second, Test Users are defined as *people and agencies that select tests, administer tests, commission test development services, or make decisions on the basis of test scores.* In the computer-based context of this dissertation Test Developers are also seen to be creators and vendors of testing software, while Test Users are understood to

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

be the institutions that utilise the computer testing software to administer computerised tests to students or Test-Takers requiring to be assessed.

In this regard the **Code** (APA, 2004) is still relevant to computerised assessments as can be understood by the following statement in the second paragraph of the **Code**:

> *The **Code** applies broadly to testing in education (admissions, educational assessment, educational diagnosis, and student placement) regardless of the mode of presentation, so it is relevant to conventional paper-and-pencil tests, computer-based tests, and performance tests.*

The **Code** is an important document for the principle of fairness in assessment, and as such has been an important guide to me and a motivation to me to do this research. I erroneously believed that the arguments in this dissertation were not being investigated by the research community because it was acceptable to society to tolerate unfair practices in assessment if an insignificant number of people were affected. The guidelines in this **Code** helped me to realise that my instincts concerning the unacceptability of even one person unfairly assessed is indeed one too many. Consequently the main question of the study is answered. It is NOT ethical to accept unfairness at all, even if only one person may be affected. This answer is important as it allowed for this study to continue answering the remaining key questions.

The APA is so concerned about ensuring assessments are conducted in a professional and fair manner that they also published a statement of several pages entitled *Rights and Responsibilities of Test Takers: Guidelines and Expectations* (APA, 1988). A quote from the third paragraph of the statement helps one understand the limited jurisdiction of the document. *This document is intended as an effort to inspire improvements in the testing process and does not have the force of law. Its orientation is to encourage positive and high quality interactions between testing professionals and test takers.* It is with the same ideal in mind that the arguments presented later in this dissertation are forwarded. It is hopefully to ensure that improvements if

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

shown to be necessary through future research studies will be made on the highlighted deficiencies in present day computer-based testing practice.

1.14.2. Test-taking Skills

The inspiration for this dissertation is grounded in the literature pertaining to test-taking skills. As a scholar and student I have always remembered to apply the recommended habits of test-taking and believe that these habits have always ensured I was able to score according to the amount of study and preparation I had invested in each test's syllabus. Since becoming a lecturer, I notice several candidates that do not apply the test-taking habits and I believe this is having a significant effect on the individual student's test results, in all their tests. The knock-on effect in terms of throughput rate will obviously be felt by institutions, particularly in South Africa with the government subsidy funding formulae heavily weighted in terms of the numbers of students graduating each year. In my reflections on the test-taking skills I am therefore amazed that this subject receives little attention in South African educational research. Be this as it may, I found some useful sources on this topic in the academic databases.

The first article on this topic focuses on test-taking skills but also alludes to the effects of applying these skills in terms of increased confidence and better test results. Glenn (2004, p. 61) states, *But, teach them (test-taking skills) to students, and watch their confidence and test scores soar.* This statement infers that students neglecting to apply the test-taking habits are indeed scoring lower than they should. Also, the self-image of each candidate and his/her self-confidence is improved. The positive affects of practising good test-taking habits/skills are evident and the spin-offs in terms of institutional throughput rates are obvious.

So what are these habits or skills? Glenn (2004) has a *During the test* list of the skills to be taught to students. The list (for during the test) from Robert Glenn follows:

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

1. *Read the entire test first.*
2. *Read the questions carefully.*
3. ***Answer the easiest questions first.***
4. *When time is a factor, don't stay on a question or problem you are unable to answer.*
5. *If you don't know an answer, skip the question and come back to it.*
6. *Read the directions for the test carefully.*
7. *Read the entire question carefully.*
8. *Complete the test before going through it a second time.*
9. *Write helpful notes in the test margins.*
10. *Go with your first thoughts when answering questions.*
11. *Put down an answer for everything.*
12. *Use the test as a source for hints.*
13. *Write legibly.*

The rationale for doing the first two habits on the list is to enable students to perform the most important habit, in my opinion, occurring third on the list, which is to plan the sequence in which to answer tests such that the student begins with their best mastered section or "easiest" question as seen from the perspective of the student. In computer-based test assessments this sequencing of questions is not controlled by the test-taker. For security reasons and to minimise cheating (Pain & Le Heron, 2003) the computer-based testing software has an algorithm that randomly sequences and presents the test questions to the students. Test security is enhanced but for students the items are not presented in an ideal sequence of easy items first, that become progressively harder. In my opinion, this may be having an adverse affect on student performance, as this important habit is one of several that allow us to *watch their confidence and test scores soar* (Glenn, 2004). The arguments presented in Chapters 2 and 3 will build on this theme.

One school in America views test-taking skills in such high regard that they have recruited the parents in a novel manner to ensure children learn and practise the skills. The Armstrong Middle School students in Starkville,

created using BCL easyPDF Printer Driver
Click here to purchase a license to remove this image

Mississippi reports on the use of parent evenings to make parents aware of the importance of test-taking skills, and on how the evenings are impacting on the parents and children of the school ("Use parent nights to improve student test-taking skills," 2004). The report does not specifically mention computer-based testing, but test-taking skills, if practised, become useful habits that can be used in all modalities of testing. Students with good test-taking habits will adapt those habits to attempt to overcome the randomisation of items by taking care to navigate back and forth through the items to do the first three habits listed by Glenn (2004). Perhaps this too can be ascertained in future studies, i.e. in computer-based assessments, are test-takers first reading all the items carefully, before navigating to their easier items and answering these first rather than attempting the items in the sequence they are randomly presented with.

Although there are several sources on test-taking skills, I will only be citing one more as the others just repeat the repertoire of recommended habits required to ensure test-takers can maximise their ability to recall studied material for the purpose of gaining the test score they deserve. This ability is stressed in a concluding statement of a book titled *Pass Your Exams Easily* (Gibson, Hoole, & Passchier, 1989) and I quote, *Remember, you can't recall information that has not been put into the system. That means you have to concentrate on learning your work and* ***revising it regularly*** *before you can engage your brain.* Chapter 11, entitled *Exam Tips* has several statements that support the arguments presented in this dissertation, and I will list these below (Gibson et al., 1989):

> *Many students have good study techniques and have planned a sound study programme, but fail to prepare themselves for the actual exam.*
>
> *Panic creates blockages in the brain's communication system and hinders memory and clear thinking.*
>
> *Read through the* ***entire*** *paper to see what is expected.*

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

Do the easy questions first. *This also builds confidence.*

The statements above refer to the students that neglect to ensure the actual exam experience will be stress free by for example arriving early, at the correct venue, practising a relaxation technique to calm any nerves, and knowing and applying the test taking habits on commencement of the test. In other words preparation for the assessment event itself is inadequate. The last two statements above correlate with the habits listed by Glenn (2004) regarding the need to plan the sequence in which questions must be attempted. I must stress that I strongly agree with Gibson et al. (1989) as my personal experience concurs that beginning with the easier questions first will indeed build confidence as the assessment progresses. This build up of confidence is vital if students are to remain calm and maintain cognitive thought processing levels (Dutke & Stöber, 2001) for the tougher questions still to be attempted, and to avoid test anxiety for the entire duration of the test.

1.14.3. Test Anxiety

So where does computer-based testing fit into this scenario? Well, modern testing software now allows students a degree of control, in that they allow the student to read all the questions or return to previous questions and edit previous answers (Ricketts & Wilks, 2002). This goes a way to alleviate test-anxiety for learners, but not far enough. Test anxiety for most authors of computer articles deals with the anxiety associated with strange software/hardware and learning to use it sufficiently to pass the assessment. This aspect of computer-based testing test anxiety is valid and justified. However, the issue of ensuring randomised test questions beginning with easy questions becoming gradually tougher is not really considered. The algorithms are not programmed to ensure that good test-setting practice is used. I argue that some students are being unfairly affected by this randomisation algorithm as they are “unlucky” enough to get several more difficult questions first and the increased test anxiety impacts on their

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

cognitive processing powers (Dutke & Stöber, 2001) and test anxiety is then worsened. The easier questions arrive too late and a fail grade results for a student that under different circumstances could have performed better with those same questions in that very test venue at that very time.

Tobias, (1985) hypothesises that test anxiety reduces the cognitive ability required to solve problems, thus leading to poor results. On the other hand, a student with good test-taking skills needs less cognitive capacity to spend on the physical elements of the test, and can therefore concentrate more on recall of the actual learning content.

Bierenbaum (2007) adds to Tobias's work and identifies a perceived alignment between instruction and assessment. She argues that students come to expect a certain style of test, and would answer in a certain way. Deviations from such expectation would 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 in so doing may be seen as unfairly disadvantaging those candidates.

Computer anxiety and computer test-taking anxiety are worth mentioning at this stage. The former is a general anxiety to using computers not necessarily an anxiety for taking tests on a computer as is implied by the latter. Computer anxious persons, not normally prone to test anxiety, may suffer from computer test-taking anxiety. Sufficient practice with the testing software on a computer may ensure the anxiety is suppressed. Such persons may remain computer anxious in other contexts, but once confidently able to use the computer testing software the tendency to show signs of test-taking anxiety may be overcome. However, a person confident in his/her ability to utilise computers but who is prone to test anxiety, may remain a test anxious test-taker no matter the modality of testing. This study is mainly concerned with test anxiety and not specifically with computer test-taking anxiety though it is conceded that computer test-taking anxiety may be a contributing factor

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

towards increased test anxiety presented by those prone to test anxiety in paper-and-pencil administered tests. 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.*

During any assessment it is almost certain that anxiety levels are likely to increase when the test-taker attempts a question and realises he/she is unable to adequately answer it. By not attempting the easiest questions first it is likely that anxiety levels will increase. If a student attempts a test in the given or set sequence, and the set test commences with the toughest questions, the test-taker will struggle, panic, and become extremely anxious. In my experience with the top student, mentioned earlier in the background to this study, this is exactly what occurred yet my top student was not considered to be an individual prone to experiencing test anxiety. With increased test-anxiety comes a decrease in cognitive processing power that in turn limits the chances of a student to perform adequately in the easier questions. Supon (2004) discusses five strategies teachers may use to assist students known to suffer from abnormally high levels of test-anxiety. She lists the need to teach test-taking skills as one of several test-wise guidelines to prepare students for tests. In her conclusion she writes that, *providing purposeful learning experiences and test-wise guidelines help students obtain maximum performance.* Clearly, test anxiety adversely affects student performance.

This study seeks to determine the extent of the problem in the computer-based test modality where the randomising of test questions negates important test-taking skills to the possible detriment of some students.

1.14.4. Equivalence Across Modalities

The **Code** (APA, 2004) has already been shown to be concerned with all modalities of testing, hence, it is not surprising that this **Code** has been used

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

somewhat as a benchmark in research seeking to investigate the degree of equivalence across modalities.

One article examining research into the equivalence of computer-based and paper-and-pencil tests (Bugbee Jr, 1996) is particularly helpful to the reader wanting to know more about the topic. The article considers previous reviews of research into the topic, the standards and guidelines for computer-based testing, a few individual studies are mentioned, and so are studies about computerised testing over time. The conclusions of the review are particularly useful and will be repeated below:

1. *Computer-based and paper-and-pencil tests can be equivalent, but it is the responsibility of the test developer to show that they are. There is NO inherent equivalence between these two forms of test administration. Looking the same does not make them the same.*
2. *Equivalence of tests is established by meeting one of the stringent criteria:*
    a. *Equal means and distributions for alternate test forms*
    b. *Equal means and distributions, reliabilities, and correlations with criterion (validity) variables for interchangeable test scores*
    c. *This equivalence can be established by comparisons of actual scores or resealed scores.*
3. *The use of computers affects testing.*
4. *Special considerations are necessary when computers are used in testing.*
5. *Users must know psychometrics and have at least a basic understanding of computer applications to effectively utilise and interpret computer-based tests (Bugbee Jr, 1996).*

In the view of Bugbee Jr (1996), the advantages of computerised testing to the test user, and even the test-taker are worth the extra costs and efforts needed to develop and validate this test modality. Some of the advantages of computerised tests are improved test security (Pain & Le Heron, 2003) and uniformity of item delivery, the instant feedback (Zenisky & Sireci, 2002) (popular with test-takers but also with faculty that now spend less time

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

marking tests), the ability to ensure item quality due to the computed psychometric data generated by the testing software, and results can immediately be used to update test result databases, saving hours of administration time keying in individual student test scores.

In the “Ancient Literature Reviews” section cited by Bugbee Jr (1996) the most relevant is the citation of a 1989 study by Wise and Plake in which they recommended computer-based tests needed to include the following three features inherent in paper-based tests:

1. *Letting test takers skip questions and answer them later,*
2. *Letting them review previously completed questions, and*
3. *Allowing them to change answers.*

Although modern testing software now includes the above three features it is important to understand that initially computer-based tests did not have these features, and that this together with the monochrome screens and screen resolutions of the time undoubtedly had an influence on the attractiveness or lack thereof to test users and test-takers from the eighties and early nineties of the twentieth century. Reading of such studies should be done with these factors in mind that contextualise the findings and associated conclusions.

A more recent study (Russell et al., 2003) examined the benefits of converting to a computer-based administered test from paper-based and particularly how validity may be impacted upon. The ability to skip, review, and even change answered items is again mentioned as an important requirement of computer-based tests, and several studies including Wise and Plake’s 1989 study are cited. Another factor to be considered is *item layout and presentation of graphics.* The modal effects studied looked at the need for multiscreen, graphic, or complex displays, all influenced by size of screen, font size, and resolution of graphics, and how these tended to negatively affect performance in computerised tests. Russell et al. (2003), note that more research is required into this modal effect, citing the findings of a 1995 study by Kolen and Brennan, that argued that modal effects are also *dependent on the particular testing programme*. According to Russell et al. (2003), a third

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

research area is investigating comfort and familiarity with computers and the correlation with performance in computerised testing contexts. Strong correlations were found, but these authors also conceded that those studies took place before computers became a common addition to home appliances.

The study by Russell and Plati (2002), shows that in the assessment of writing skills, that forth grade students perform better on computers than using paper-and-pencil, and ascribe this to their comfort with using computers at home.

Although this study is more concerned with the multiple-choice question format, the comfort and abilities of today's youth with technology is an important variable that must somewhat date past studies. Due to the prevalence of computers in households, comfort and confidence with this modality is changing. Technology has also improved the capabilities of hardware; screens are larger, with colour monitors, and excellent screen resolutions. All this may mean that past studies may need to be repeated to ensure that they are still valid for the computer savvy test-taker of today using far superior computer hardware and software.

1.14.5. Error Variance

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 pilot 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. This means that statistically significant correlations are unlikely to be realised in this pilot study. However, there is a need for further research

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

into the effects highlighted by this study. Care must be taken in such research to ensure that the element of chance is removed so that statistically significant correlations may be possible.

1.14.6. Psychometrics

Difficulty Index and Discrimination Index are two of the psychometric indices that are readily available in computer-based tests. The psychometrics referred to in this study are from Classical Test Theory psychometrics (Reise & Henson, 2003). 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).

This study is limited to computer-based testing and hence the use of psychometric indices from classical test theory is adequate, however, if the test mode for this study was computer-adaptive testing then Item Response Theory psychometric indices would be required (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 some 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 (Alessi & Trollip, 2001).

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

1.14.7. Cognitive Loading

Cognitive workload is an important factor in testing. Literature indicates that *High levels of test anxiety are known to cause decrements in cognitive performance* (Hembree, 1988) as cited in Dutke and Stöber, (2001, p. 381). The thinking is that test anxious students are less able to perform due to negative thoughts adding to their cognitive workload and as such clogging the working memory with worries and concerns instead of with the information needed to think through a test item. Working memory is needed to cope with the item being attempted, information presented in the item stem needs to be processed and compared with the options such that the key can be correctly identified and the distracters avoided. Conceptually working memory consists of three active storage subsystems (Baddeley, 1992; Baddeley, 2003). First storage for audible information, second storage for visible information, and the third Baddeley calls the *central executive* which processes and stores the information between the three subsystems. Dutke & Stöber (2001) mention three studies that show test anxious students have a lowered cognitive performance because of how the storage and processing subsystems of the working memory are affected during assessments. The randomised sequence of test items in computer-based test assessments may be adding to the demands expected of the working memory and hence the importance of the arguments presented in this paper.

Dutke & Stöber (2001) cite three studies that found test anxious students to be adversely affected by increased cognitive loading, however, it is interesting that in their own study they found that, if anything, test anxious students were advantaged by increased demands on the working memory. In their study they differentiated task complexity into coordinative and sequential complexities. Dutke & Stöber's (2001, p383) study investigated test anxiety and sequential complexity by consideration of *memory performance in a task with high coordinative complexity under low and high sequential demands*. Coordinative complexity may be seen as the need to store the results of processed steps for use in subsequent processing of information. Sequential complexity is the storage of steps of processed information that are not

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

required in the next step of processing and may in effect be replaced without affecting the present processing stage. According to Dutke & Stöber (2001), the three studies show that increased coordinative complexity demands on working memory adversely affect test anxious students in assessments. Hence they also investigated the potential for test anxiety to occur in students when taxing the working memory with sequentially complex tasks. In contrast to the findings of the other studies, they found that high sequential complexity tasks tend to benefit test anxious students but recommend further research needs to be done in this regard.

The relevance to this paper is due to the task complexity introduced into computer-based test assessments through the randomisation of test items. Both coordinative and sequential complexities are introduced. Coordinative complexity is introduced through different items from a particular portion of the syllabus requiring the processed results of previous items to be recalled to assist in processing information in subsequent items. Sequential complexity is introduced through items from different independent sections of the syllabus allowing working memory contents to be updated and replaced for processing of later items. The concern is that the effects of the task complexities are not being considered by computer-based test developers and users, and that this may be detrimental to test-takers.

1.14.8. Primacy and Recency

Linked to the concepts of coordinative and sequential complexities and their effect on cognitive memory workload are the concepts of primacy and recency effects (Bemelmans, Wolters, Zwinderman, ten Berge, & Goekoop, 2002). In the modal model of primacy and recency, memory and recall define a short-term store (STS) and a long-term store (LTS) (Bemelmans et al., 2002). The STS can be visualised as a fixed-capacity memory buffer. **Primacy** states that the earlier items to be recalled are in the buffer longest and hence this ensures their transferral to the LTS aiding recall. **Recency** states that the last

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

items needed to be recalled are still in the buffer at the time the test is administered and so are accurately recalled (Bemelmans et al., 2002).

Primacy effects may be relevant to this paper's argument because the initial items in the test from a particular portion of the syllabus being tested will be in the STS buffer the longest and are more easily recalled.

Recency effects may be relevant to this paper's argument because the section of the syllabus relating to the last or present item being attempted by the candidate being freshest in the STS may be recalled more readily for subsequent items. However, if the subsequent item comes from a different section of the syllabus, the STS buffer must now be replaced with newer content relevant to the new item. Recency as an effect is now attenuated and now less likely to occur and hence an increased cognitive workload occurs for each candidate due to this constant changing randomly between sections of the syllabus because of randomising the item sequence in computer-based testing contexts. This may be expected to hinder test-takers in the computer-based testing modality.

Support for the relevance of the attenuation of the recency effect to this paper may be found in a paper by Talmi and Goshen-Gottstein (2006). These authors cite an earlier paper in which they found that randomised test items in a multiple-probe recognition procedure will result in the attenuation and possible elimination of the recency effect. Multiple-probe procedures test for items to be recognised from a list by randomly presenting those items among distractor items and items from earlier lists which are also distractors for the present list. The single-probe approach differs from the multiple-probe in that a study list of items to be recognised is followed by a test containing only items to be recognised and distractor items that did not feature in earlier recognition lists (Talmi & Goshen-Gottstein, 2006). One should note though that in this cited study items are not multiple-choice questions but are entities on a list to be recalled. However, the multiple-probe recognition procedure is synonymous with the many different sections of the test syllabus of this study. The randomised sequence, in which test items draw from these different

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

sections, may therefore attenuate the recency effect, and impede effective and efficient recall for the test-taker. The suggestion to use a single-probe approach (Talmi & Goshen-Gottstein, 2006) is reported to avoid the "test order confound". In the context of computer-based testing one may infer that a single-probe approach can be obtained by ensuring that multiple-choice items are grouped according to the section of the syllabus from which they are sourced. The logical question now to ask is: Can grouping together of multiple-choice items from a section of the test syllabus avoid a "test order confound" by modelling a single-probe approach for each section being assessed?

To sum up the overall affect for test-takers is the difficulty to practise good test taking habits that may affect them emotionally (Mealey & Host, 1992; Schutz, Davis, & Schwanenflugel, 2002). Coupled with this the sequential and coordinative task complexities introduced may affect their cognitive functionality (Chapell et al., 2005), and both of these issues are a direct result of the randomisation of test items in computer-based test assessments. In addition, memory and recall are key faculties a test-taker must rely on in testing contexts, yet the effect of randomised test items on recall through attenuation of the primacy and recency effects is a concern for me. In high stakes testing scenarios the effects will be more pronounced (Hancock, 2001; Keogh & French, 2001) and yet these are the situations in which fairness and an ability to accurately and efficiently recall study material is even more critical due to the impact of the test performances on the futures of individual test-takers.

1.15. Research Methodology

1.15.1. The Null Hypothesis

This is an ideographic study as even if only one person of the sample group is affected, this is enough to advocate for future research studies to be done to

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

investigate how preventative action is taken to ensure this is not continued in future assessments.

This study sought to show that the performance of students is significantly affected if they are presented with difficult questions first during a computer-based test.

However, in order to conduct this pilot study, I propose the following null hypothesis: "There is no significant difference between the performances of students presented difficult questions at the start of a randomised multiple-choice question computer-based test, compared to those presented the easy questions first."

1.15.2. The Pilot Study

The sample group consists of 103 Veterinary students in the subject coded *BHP 470*. The subject is assessed by way of computer-based tests and a database of questions is used. The software used to administer the tests is *Questionmark Version 3.2*. This sample is chosen due to the relatively large number of students, and because the data set consists of four computer-based test assessments conducted in 2004. The four sets of assessment data available for this group of students allows me to analyse the data for correlations that may prove/disprove the stated null hypothesis, without the need to access confidential student records. By this I assume for the purpose of this pilot study that each student's performance is obtained by the mean of that individual student's scores obtained for the four tests in this data set. The correlation between a deviation in any student's performance and the sequencing of the difficulty of the questions for that student allows me to make findings on the plausibility of the null hypothesis.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

1.15.3. Cleaning of the Data

The raw data for each of the four computer-based test assessments was processed to allow it to be statistically analysed by the Statistics Department of Pretoria University. The raw data consisted of four text files and one *personal document file* for each of the four assessments. The following coding was used to help organise the data:

- The assessments were numbered "1" through "5". Only four assessments were analysed but data for five tests was given to me. The fourth test was disrupted by a technical problem that occurred minutes after the assessment began. The entire set of questions for this test was discarded, and a new fifth test was created from the question databank. The test was rescheduled and taken by the sample at a later date. The text files for the fourth test have not been analysed but they were still numbered so that there would be no confusion for scholars that wanted to replicate my study, concerning what was done to the extra test data that was obviously unusable due to the test not being completed.
- The four text files for each assessment were labelled "a" through "d" and represented the following useful information:
    a. This is the "List Report.txt" file containing the Final Score listed next to each test-taker's student number.
    b. This is the "Summary Report.txt" file containing the basic statistical indices for the particular assessment.
    c. This is the "Full Report.txt" file containing the response to each item for each student, in the particular sequence the questions were presented to each student, and the score obtained for each item.
    d. This is the "Question Analysis Report.txt" containing the master sequence of the questions, and the sample statistics relevant to each item.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

- The fifth file for each assessment was a *personal document file* script and as there was only one per assessment it did not require a letter as a post-label.
- The files all were given my surname as a pre-label, i.e. the 2nd text file of the 3rd assessment was coded/named, “marks3b”, and the 2nd text file of the 5th assessment was coded, “marks5b”.

The above data sets are typical of *QuestionMark 3.2*. All the psychometric data information is available but in an illogical mix of text files, making statistical analysis quite difficult. The cleaning process for the statistical comparison required by this study necessitated the actual answer for each student, and the sequence that the set of questions were presented to each student. This information was spread across the four text files “a” through “d”, so the task was made more difficult. Advanced spreadsheet skills were required to sort, sift and consolidate the data into a format which could be useful for the statistics analysis.

1.15.4. Limitations of this Study

The study is based on existing computer-based test data. However, in order to research Item Randomisation Effect thoroughly, data should be collected according to a rigorous research design plan. This was not possible for the pilot study but is mentioned as a limitation.

The requirement to find the largest sample group in order to improve the likelihood of obtaining statistically significant correlations meant that a sample consisting of veterinary students was selected, as this was the largest sample group available. The randomisation of multiple-choice items in a computer-based test is likely to affect test anxious test-takers more than those candidates that are mature test-takers with good test-taking skills. Veterinary students are the crème of the crop of school-leavers, and as such can be regarded as mature test-takers with adequate skills for taking tests. A better sample should consist of first year students in a general ability group of

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

students, not an elite group as veterinary students most certainly are. However, as is reported in Chapters 2 and 3, some remarkable occurrences are still noticeable in spite of the sample consisting of veterinary students.

1.15.5. Constraints

This study is an ex post facto study based on existing data. The research is a dissertation of limited scope, hence the decision to use existing data and treat this as a pilot study to inform possible future research in this area.

1.15.6. Ethics

Ethical clearance was deemed to be unnecessary for this pilot study, as no students were used, just their existing computer-based test data was analysed. The data was obtained a year after the students completed the course, and as such had no influence on their result for the course. The students' normal performance level was based on their performance for the four tests in the course, so access to confidential student academic records was also not required. Permission to use the data was forthcoming from the Veterinary Sciences Faculty of Pretoria University.

Future research designs may require ethical clearance as sample group student results may be affected, and access to academic records may be preferred over the method used in this study to define normal performance for each test-taker.

1.16. Structure of the Thesis

The thesis is built around the arguments contained in two articles that have been submitted for publication in the Information Technology and Society journal. Each of these articles makes up a chapter and become Chapters 2 and 3 of this thesis. The articles are co-authored by my supervisor; however,

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

for the purpose of this dissertation the style used in Chapter 2 and Chapter 3 is converted into a first person singular style. The introduction chapter, Chapter 1, and the concluding chapter, Chapter 4, are written around the two articles making up Chapter 2 and Chapter 3.

The attentive reader may notice that each article tends to duplicate parts of the literature review given in Chapter 1. This is because each article required its own literature review to enable each to stand alone, and be readable to persons that had not yet read this dissertation, or even the contents of the other article. Your understanding in this regard is appreciated.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

2. Chapter 2: Russian Roulette in Assessment?

2.1 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 disadvantage certain 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 where instances were found where students may either have been unfairly advantaged or disadvantaged by being given a series of easy, and/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 I have named, "The Item Randomisation Effect".

2.2 Keywords

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

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

2.3 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 (constructor) wishing to ensure that items progressively become tougher, and although navigation through computer-based tests is possible it is certainly an inconvenience to test-takers wanting 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 each test-taker is administered the test items in a different sequence to 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 to play in high stakes testing contexts. Russian Roulette does not cause injury to all the participants, yet is deadly to the one that 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

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

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 to improve student test-taking skills," 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 that have good test-taking skills with regards to paper-based tests are still going to 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, causing increased anxiety. Increased anxiety at any stage during the test for whatever reason is likely to have a negative effect on that persons' 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 on 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 disadvantage any of the test-takers in any way?"

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

The following sub-questions guide the study:

What constitutes “normal performance” for each student in the sample?

Has any candidate’s performance differed significantly from his/her “normal performance”?

Have such candidates been presented with consecutive randomly sequenced difficult items?

May this deviation be attributed to the randomisation of items presented to the student?

‘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 of a list occurring at the end of the list. Studies of these effects (Bemelmans et al., 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 are different multiple-choice questions, not a list consisting of word (items) and or number (items) to be recalled in the correct sequence. The items or questions in this study are randomised so that test-takers get administered the same test but for a few test-takers the randomised sequence of items may subtly cause some candidates to be seemingly disadvantaged in comparison to the rest of the candidates taking the test.

For the purpose of this article/chapter I felt that primacy and recency effects on memory and cognition should be excluded from the present discussion. This chapter focuses on the Item Randomisation Effect, and for the sake of clarity, primacy and recency are discussed separately in Chapter 3.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

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, e.g. entrance examinations to tertiary institutions, promotion or retention of children in schools, or selection of potential candidates for a particular job vacancy (Russell et al., 2003; Zenisky & Sireci, 2002). Consider what it would be like to be the test-taker administered the items 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. It is hoped that this study might justify further research into the Item Randomisation Effect to ensure that all test-takers are always fairly administered tests in future.

2.4 Literature Survey

Tobias, (1985) hypothesises that test anxiety reduces the cognitive ability required to solve problems, thus leading to poor results. On the other hand, a student with good test-taking skills needs less cognitive capacity to spend on the physical elements of the test, and could therefore concentrate more on recall of the actual learning content. Bierenbaum (2007) adds to Tobias's work and identifies a perceived alignment between instruction and assessment – she argues that students come to expect a certain style of test, and would answer in a certain way. Deviations from such expectation would 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 in so doing 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

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

guides, and they are regarded as the stakeholders specifically tasked with ensuring the guides 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 (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 impacts negatively on test results, while navigational control enhances test performance. No literature could be found that comments specifically on 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, as it distracts from the holistic nature of a test.

Pain & Le Heron, (2003) report that randomised sequencing of test items has been successful in preventing cheating in computer-based tests. In one of the scenarios these authors 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 no students could cheat, as each student answered a different set of questions. However, the fairness of such a solution is questionable, and they returned to rather allowing random sequencing of a set of test questions, such that all test-takers essentially took the same test. However, the potential of unfairly presenting sequences of difficult items early in the test is not considered by these authors.

Glenn (2004, p62) advises test takers to: *Answer the easiest questions first. Completing the sure-thing questions first boosts student confidence from the outset*. Literature does not say what happens to student confidence or anxiety if this important test-taking skill is ignored. I would like to deduce that the opposite effect will occur, student confidence wanes and test anxiety

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

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 increased tension and anxiety ("Reduce Test Anxiety to Improve Student Performance," 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 that student performance was adversely affected by it. Computer anxiety is prevalent in persons that seldom use computers in their daily lives, hence their nervousness when using computers. This computer anxiety compounds the natural anxiety due to 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 identified 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 guides allow for scores to be adjusted such that the adjusted score represents a fair representation of the test-takers when compared with persons taking tests in other test modes (Russell et al., 2003).

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

However, the item randomisation test mode effect identified by this study has not as 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 to improve student test-taking skills," 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 readily shows that by allowing candidates more control, the disadvantage could be somewhat negated. However, it now becomes obvious that students need extra time to navigate back and forth through the test items as compared to students sitting for the same test in a paper-based mode.

The mode effect described in my 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 has not published any studies in this regard.

2.5. Method

2.5.1 Overview

In this study I investigate the performance of 103 veterinary science students 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

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

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 type 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 were manipulated in any way. The data was simply collected from the computer-based testing department the year after the tests had been administered as this was believed to be sufficient for the purposes of this pilot study. The data was then analysed and the normal performance of each student was established as described below. The test data was processed until the relevant variables were ready for analysis. Next the students that deviated significantly from their normal performance were flagged, and their test experience was then analysed in detail. The graphic 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 performance of the test.

2.5.2. 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. A significant deviation in the performance of that candidate needs to be reported. One obvious record of normal performance can be obtained from each student's academic record, and the average obtained thus far in his/her academic career. However, this is not satisfactory as deviations above and below the average is normal, according to the individual student's aptitude and motivation levels for different courses. I decided rather to find a sample of students that had sat for more than two computer-based test assessments in one year-long course. The average obtained for each candidate for the particular course would, for

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

the purpose of this study, be considered as the normal performance of the student.

2.5.3. Defining Difficult Items for this Study

Difficulty Index and 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 some 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.

2.5.4. Sample Data Selection for this Study

The Computer-based testing section of the University of Pretoria assisted me 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 for four computer-based test assessments. *Questionmark* 3.2 was

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

the software used to administer the tests. The data available for each test consisted of four text files generated by *Questionmark* after 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.

2.5.5. 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 of problems during the test that required the test be 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 of the five 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 other potentially useful information. The process was repeated for each of the four sets of test data. The data was now ready for analysis.

2.6. Discussion from Findings

Detailed analysis of the particular students that had test scores varying by at least fifteen percent seems to show correlations between the sequences of item difficulty and the effect on student performance for the test. Literature

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

indicates that beginning with easy items in a test is one habit likely to have a positive affect on student grades (Glenn, 2004). In this study, students that scored significantly higher than normal in Test 1 tended to be presented with the easy items early in the test. Conversely, those presented with the difficult items early tended to perform significantly below their normal levels.

The figures that follow are the most striking examples, one from each of the tests of the trends that become apparent, when one studies the data pertaining to the candidates that 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 per Item,
2. The horizontal line labelled Average Difficulty is also the average scored for the particular test,
3. Diamonds 10% above the Average Difficulty can be seen as “Easy”,
4. Diamonds 10% below the Average Difficulty line can be seen to be “Difficult”,
5. Diamonds within 10% of the Average difficulty can be seen as “Average”,
6. The candidate’s deviation from normal is indicated under the chart title.

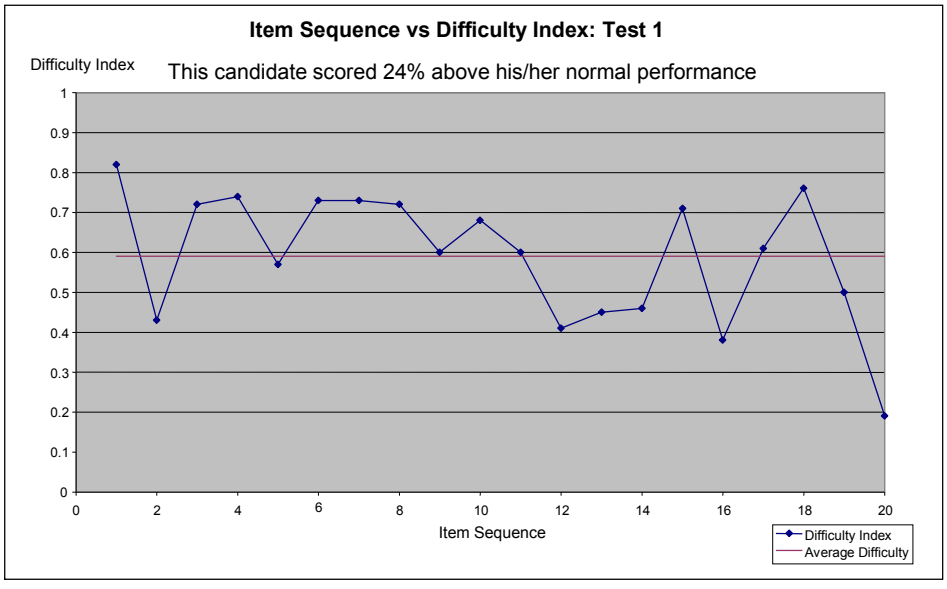

Figure 4: Example from Test 1

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

Figure 4 (previous page) shows an example of a candidate that scored twenty-four percent *above* his/her normal level. 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, supporting the literature pertaining to attempting easy items first (Glenn, 2004; Supon, 2004).

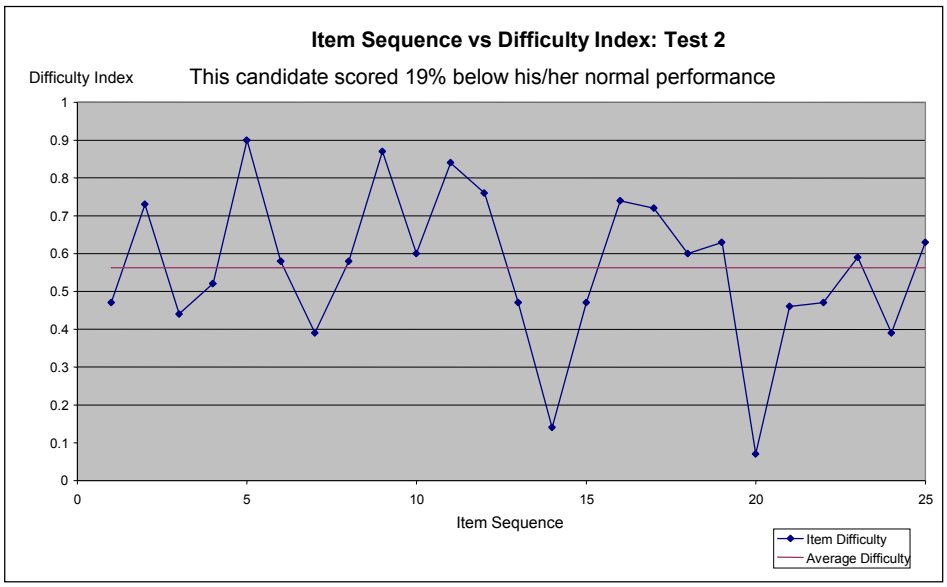

Figure 5: Example from Test 2

A closer look at Figure 5 shows an example of a candidate that scored nineteen percent below his/her normal level. For the first eight items, two were easy, two difficult and four of average difficulty. The ideal would be easy items in the beginning, average items in the middle, and difficult items at the end. It is also important to remember that the assumptions that 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.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

In Figure 6 the candidate obtained fifteen percent below his/her normal level. The first third of the test or the first sixteen items indicate that four were easy, six difficult, and six of average difficulty. Most of these are average to difficult, which ideally should be the case for the middle third of the test. Did this random sequence of test items cause increased anxiety?

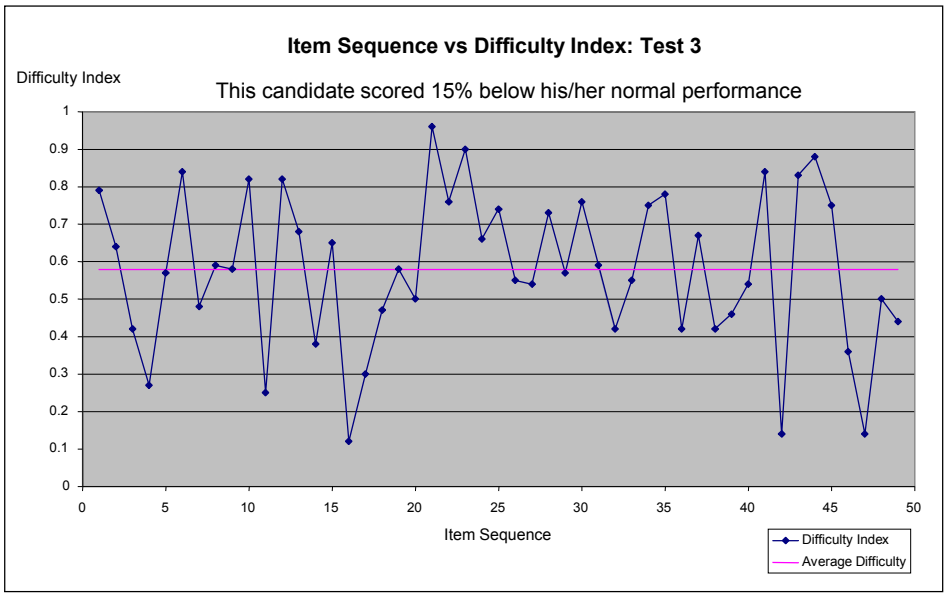

Figure 6: Example from Test 3

Figure 7, (next page) shows a candidate from the Test 5 that scored fifteen percent below his/her normal performance in the four tests. For the first sixteen items, three were easy, four difficult, and nine of average difficulty. Again this is showing the initial third of the test to be of average difficulty level, which in the ideal should be the middle portion of the test. Even closer inspection shows only six items above the average line, and ten items were on the difficult side below the average line for the first sixteen items. The initial items presented to this candidate are clearly tending to be the more difficult items for this test, and this indicates to us that this student too could be a victim of unfair assessment by being randomly presented with difficult items early in this test.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

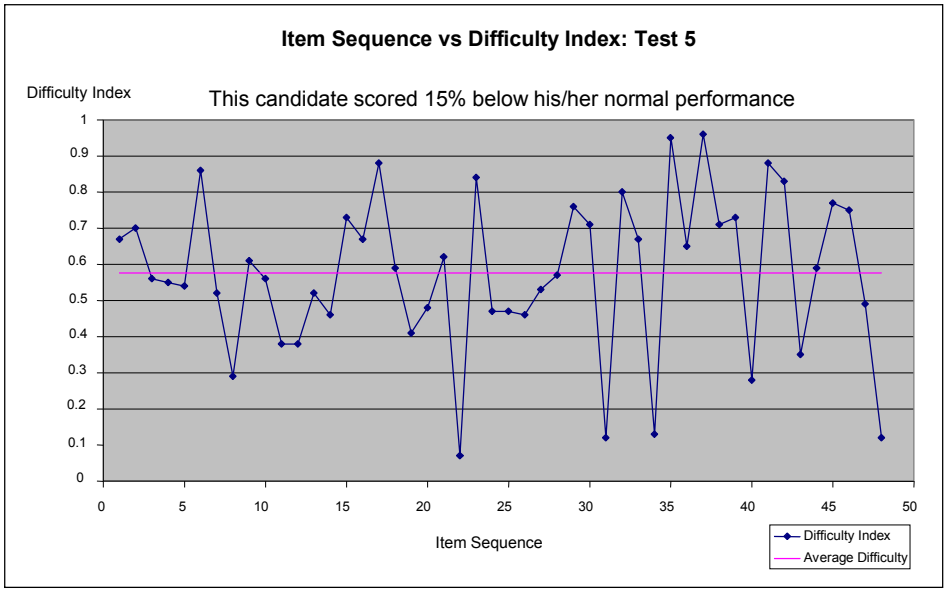

Figure 7: Example from Test 5

2.6.1. Limitations

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 type 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: Scores Varying by at Least 15% from Normal

	Test 1	Test2	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%**

Table 1 above illustrates the problem with analysing the data. For each of the tests, fewer than ten percent of the sample scores vary by at least fifteen percent from their normal performance, calculated as explained earlier in this paper. In fact, the percentage drops below five percent for the last two tests.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

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 investigation of potential correlations.

This is a pilot study using pre-existing data with an ideographic concern for the performance of individuals taking the tests. As a result formal statistical analysis is unlikely to show statistically significant correlations. My initial design for an experimental type of research study was considered to be beyond the scope of this dissertation, but I believe future research should be designed to ensure the data is nomothetic. Perhaps the test items could be administered to the control group in the ideal sequence (items become progressively more difficult) while the experimental group are administered the test items in the worst possible sequence (items become easier). Ethically, the experimental group may be disadvantaged by this design, so it should be conducted with great care.

The above limitations are explained further in terms of construct validity in Bugbee Jr (1996), with regard 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 ensure all candidates are presented 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

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

be taken to account for those candidates with good test-taking skills and low levels of test anxiety that will probably attempt the items in a progressively more difficult sequence influencing the reliability of the study.

2.7. Conclusions and Recommendations

Despite the limitations the findings of this study are noteworthy because a potentially unfair testing practice has been identified and verbalised. Also a possible gap in the literature can now be filled as researchers investigate this test mode effect that in this study is referred to as “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 that 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, and it is the responsibility of test users to ensure developers are made aware of this test mode effect. This unintended Russian Roulette in assessment is a violation of the principle of fairness for each test-taker as required by the internationally recognised **Code** (APA, 2004).

Assuming this mode effect is found to cause some students to be disadvantaged in computer-based test assessments, I would recommend that software vendors modify 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, labelled and stored in the item database. Randomising algorithms can now present items to students randomly but progressively increasing the difficulty of the items. Randomisation ensures test security, and also allows items to become more difficult as the test items are presented to each test-taker thereby preventing possible Item Randomisation Effect.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

3. Chapter 3: Rollercoaster Ride in Assessment?

3.1. Abstract

Research into effects pertaining to non-adaptive computer-based testing is more concerned with the performance of the test, than with the performance of individual test-takers. Surely a test is meant to also serve the interests of every honest test-taker. This article builds on the Item Randomisation Effect coined by Marks, Cronje, & Mostert (in press) and is a result of reflections and discussions around the graphical results presented in that study. Two related issues are to be found in this article.

First, the need for test-takers to adjust to drastic changes in item difficulty from one item to the next results in an emotional fluctuation between relief and panic throughout the test.

Second, the possible increase in cognitive workload caused by a random dispersion of items from any one section of a syllabus may be such that working memory is excessively taxed. For both it is the test anxious candidate that may be more significantly affected by this item randomisation in computerised testing. Graphic evidence showing the drastic changes in difficulty of successive items is presented as a pictorial comment on the practice of item randomisation. I believe further research into the Item Randomisation Effect is required.

3.2. Keywords

Computer-based tests, Fairness, Cognitive workload, Item Randomisation Effect, Anxiety.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

3.3. Introduction

Fairness in testing is an important ideal to be strived for by all test developers and test users. In this regard guidelines have been compiled and revised by different authorities, perhaps the most renowned is the American Psychological Association (APA, 2004). In this code **Test Developers** are defined as those vendors that create computer-based testing software to be used in testing venues for administration of tests via computers to test-takers. **Test Users** are defined as those institutions that use computer-based testing software to administer tests to test-takers, and include universities, colleges, and potential employers. **Test-takers** are defined as those persons that are being tested to obtain a measure of their abilities and potential suitability for future career or scholastic promotion. The principle of "fairness" is of such importance that the APA has published a document to fully inform all testing stakeholders particularly test-takers of their rights and responsibilities (APA, 1988). The perspective of this paper is to highlight potential threats to fairness in present computer-based testing practice, particularly to test-takers, that result from the need to randomise items in computer-based testing. Item randomisation in computer-based test assessments may solve the problems of cheating in computer-based test venues (Pain & Le Heron, 2003); however, they may also cause random individual test-takers to be unfairly disadvantaged. Fairness in testing is an important principle underpinning all assessments, and it is a right of all test-takers to be tested in a fair and equitable manner. Most research on test results is predominantly nomothetic because it is concerned with the performance of the test; this research tends to be ideographic because it is concerned with the performance of the individuals taking the test.

Computer-based administration of multiple-choice questions evolved out of a paper-based format. Despite the inherent problems, its advantages in terms of instant feedback, reduced marking times, convenience for students (Pain & Le Heron, 2003), and item statistics reports (Ferguson et al., 2002) mean that solutions for the inherent problems must be found (Hall, 2000). So far much

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

has been done to ensure equivalence across the paper-based compared to computer-based modalities (Bugbee Jr, 1996; Clariana & Wallace, 2002). Now test-takers have the ability to navigate through the tests, to skip items returning later to change previous answers, etc., which are now standard features of computer-based testing software. This chapter/article expands on the Item Randomisation Effect identified by Marks et al. (in press) discussed in Chapter 2. The underpinning principle of fairness in assessment for all test-takers (APA, 2004) is the justification for the perspective contained in this paper. The Item Randomisation Effect (Marks et al., in press), indicates that test-takers randomly presented with a sequence of difficult items early in a test might be unfairly caused to perform poorly in that test. This chapter will now discuss further consequences of Item Randomisation Effect. The two additional consequences are that randomly presenting items may cause, (i) an increased cognitive load on test-takers and, (ii) a fluctuation of emotions. The concern is that these additional consequences can also adversely affect the performance of some test-takers in a test, particularly those prone to test anxiety.

The Item Randomisation Effect identified by Marks et al (in press) produced graphic representations (see Figures 4 – 7 from Chapter 2) plotting Item Difficulty Index versus Item Sequence for each test-taker. Analysis of these graphic representations for test-takers that have scored significantly lower or higher than their normal performance levels in any of the computer-based tests highlights the possible existence of the two additional test mode effects that are discussed in this chapter. The graphs shown in Chapter 2 plot a line representing the difficulty and sequence of items presented to candidates in the study. In each of the four plots from Chapter 2, the line fluctuated above and below the average difficulty level regularly throughout the assessment. This paper asks whether this continual requirement to adjust to large fluctuations in item difficulty level might cause an emotional rollercoaster of relief and panic to follow on one another throughout the test.

Another effect relating to the increased cognitive load on test-takers that is introduced through the randomisation of test items has to do with the portions

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

of the syllabus that the items are assessing. Items that are truly randomly presented will certainly tend to ensure that different sections of the syllabus are presented in a random sequence too, not allowing sections of the syllabus to be presented consecutively or grouped together as is normal in paper-based tests. I argue that this adds to the cognitive workload of test-takers in computer-based tests as they are required to keep information from earlier items in their working memory as long as possible in case another related item prompts for this information later in the sequence of presented items. This is compounded by the next item that causes a similar demand on the working memory, and the next, and the next, assuming a sequence of four items from four different sections of the syllabus is presented consecutively.

This paper is a comment on present computer-based test practice, and on potential effects that need to be studied by researchers to ensure unintentional unfair testing practice is prevented in future testing events. Although the two effects described above will affect all the test-takers to some degree, it is likely to affect test anxious candidates more than those not prone to test anxiety.

3.4. Literature Survey

Randomised sequencing of multiple-choice question items in computer-based tests is an accepted practice (Pain & Le Heron, 2003; Wang, Wang, Wang, Huang, & Chen, 2004) mainly due to the ease with which neighbouring screens can be viewed by other test-takers. This randomisation of test items helps to ensure tests remain valid and reliable measures of test-takers' abilities because viewing another candidate's screen is not helpful if the item presented on that screen is different to the one being attempted by the would be cheater. However, test validity and reliability must be achieved whilst ensuring a third important criterion, fairness to all test-takers in the test.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

With the prevalence of computer-based testing, the APA has also published guides (APA, 2004) to ensure that testing must be fair across modalities too, meaning takers of the standardised test in paper formats should score the same result even if they had been administered that test via computer. It is the position of this paper that the randomisation of test items administered by computer may be causing fairness to be compromised for some test-takers across modalities. As those administered the test by computer may be experiencing emotions and cognitive loading not required of candidates in paper-based testing environments. The only literature commenting on the effect of item randomisation in computer-based tests is by Marks et al (in press). However, this paper adds to the concerns of Marks et al due to subsequent reflection on that article's findings.

A particular concern for me is the difference across test modalities in respect of test-taking skills. It is well documented that test-takers should develop certain good habits to practise and use whilst sitting for tests, and that by doing these habits their test scores will be maximised (Glenn, 2004). These habits will, in effect, allow takers to score according to their level of preparedness for the test set by the test user. It is inferred that test-takers neglecting to practise these habits in tests score significantly lower than their level of ability (Marks et al., in press) due to the resultant reduced evidence of their abilities submitted in that test. These test taking habits are particularly useful and practicable for test-takers in paper-based tests. However, as will be highlighted in this paper, I believe that the randomisation of test items makes it difficult for test-takers in computer-based test assessments to apply all of the recommended test-taking skills. The concern is that this may result in some test-takers being disadvantaged, particularly those with a tendency to test anxiousness that may not be as easily able to adapt to the variation in emotions (Schutz & Davis, 2000; Schutz et al., 2002) and/or the extra cognitive burden the random administration of test items may place on them (Baddeley, 1992; Dutke & Stöber, 2001; Keogh & French, 2001).

The specific habit that is more easily practised by test-takers in paper-based tests is the habit requiring candidates to begin the test by attempting their

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

easiest questions first (Glenn, 2004) and to leave the tougher questions for the end of the test. In paper-based tests several items are presented on each page of the test, and the test items can easily be attempted in any sequence by paging back and forth. However, presenting more than one item at a time on a screen in computerised tests causes fatigue and irritation to test-takers especially if this meant scrolling was required (Ricketts & Wilks, 2002). So it is now recommended in computer-based tests to present items one at a time and to take care to limit scrolling by test-takers as far as is practicable. The result of this together with the randomisation of items is that the convenience of attempting items in a sequence determined and desired by the test-taker is compromised. This has already been highlighted by Marks et al (in press) however another pattern has emerged, from reflection on the graphs given in Chapter 2, that is noteworthy and has resulted in the two additional effects commented on in this chapter. The pattern that emerged was a criss-crossing of the average difficulty line throughout the test. In other words, easy and difficult items presented randomly such that easy frequently followed on difficult and vice-versa for the duration of the test.

The criss-crossing causes two additional effects. Firstly emotions may be caused to fluctuate due to the relief when presented easy items contrasting with the dismay/panic when presented with difficult items. Secondly, items are not grouped according to the portions of the syllabus but logically related items are randomly dispersed throughout the test. This is not surprising as the randomising algorithm is supposed to randomly present test items to candidates thereby ensuring neighbouring test-takers are presented different items throughout the assessment. In the paper-based test mode, items are not usually randomised, as the likelihood for cheating is significantly lower. In the paper-based test mode items from different sections of the syllabus are grouped together allowing candidates to attempt all the items from a section of the syllabus before progressing to other sections. In computer-based test assessments, the random presentation of items to test-takers will almost certainly ensure a constantly changing level of difficulty, and a constantly changing section of the syllabus presented consecutively. This may cause increased cognitive workload on the test-taker with an increased likelihood of

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

test anxiety in some candidates. A veritable roller coaster ride of ups-and-downs, twists-and-turns for test-takers may result. My main concern is that the above effects are apparently not considered by test developers and test users, and that test-takers are expected to make the best of the possible unintended consequences of allowing a computer algorithm carte blanche in the random presentation of items regardless of (i) item difficulty sequence, or (ii) the sections of the syllabus from which consecutive items are sourced.

Cognitive workload is an important factor in testing. Literature indicates that *High levels of test anxiety are known to cause decrements in cognitive performance* (Hembree, 1988) as cited in Dutke and Stöber, (2001, p. 381). The thinking is that test anxious students are less able to perform due to negative thoughts adding to their cognitive workload and as such clogging the working memory with worries and concerns instead of with the information needed to think through a test item. Working memory is needed to cope with the item being attempted, information presented in the item stem needs to be processed and compared with the options such that the key can be correctly identified and the distracters avoided. Conceptually working memory consists of three active storage subsystems (Baddeley, 1992; Baddeley, 2003). First storage for audible information, second storage for visible information, and the third Baddeley calls the “central executive” which processes and stores the information between the three subsystems. Dutke & Stöber (2001) mention three studies that show test anxious students have a lowered cognitive performance because of how the storage and processing subsystems of the working memory are affected during assessments. The randomised sequence of test items in computer-based test assessments may be adding to the demands expected of the working memory and hence the importance of the arguments presented in this paper.

Dutke & Stöber (2001) cite three studies that found test anxious students to be adversely affected by increased cognitive loading, however, it is interesting that in their study they found that if anything test anxious students were advantaged by increased demands on the working memory. Dutke & Stöber’s study differentiate task complexity into coordinative and sequential

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

complexities. Dutke & Stöber's (2001, p. 383) study investigated test anxiety and sequential complexity by consideration of *memory performance in a task with high coordinative complexity under low and high sequential demands*. Coordinative complexity may be seen as the need to store the results of processed steps for use in subsequent processing of information. Sequential complexity is the storage of steps of processed information that are not required in the next step of processing and may in effect be replaced without affecting the present processing stage. According to Dutke & Stöber (2001), the three studies showed that increased coordinative complexity demands on working memory adversely affected test anxious students in assessments, hence their interest in studying the potential for test anxious students when also taxing the working memory with sequentially complex tasks. In contrast to the findings of the other studies, they found that high sequential complexity tasks tended to benefit test anxious students but recommended further research needs to be done in this regard.

The relevance to this paper is due to the task complexity introduced into computer-based test assessments through the randomisation of test items. I believe both coordinative and sequential complexities are introduced. **Coordinative complexity** is introduced through different items from a particular portion of the syllabus requiring the processed results of previous items to be recalled to assist in processing information in subsequent items. **Sequential complexity** is introduced through items from different independent sections of the syllabus allowing working memory contents to be updated and replaced for processing of later items.

The concern is that the effects of the task complexities are not being considered by computer-based test developers and users, and that this may be detrimental to test-takers.

Linked to the concepts of coordinative and sequential complexities and their affect on cognitive memory workload are the concepts of **primacy and recency** effects (Bemelmans et al., 2002). In the modal model of primacy and recency, memory and recall define a short-term store (STS) and a long-

created using BCL easyPDF Printer Driver
Click here to purchase a license to remove this image

term store (LTS) (Bemelmans et al., 2002). The STS can be visualised as a fixed-capacity memory buffer. **Primacy** states that the earlier items to be recalled are in the buffer longest and hence this ensures their transferral to the LTS aiding recall. **Recency** states that the last items needed to be recalled are still in the buffer at the time the test is administered and so are accurately recalled (Bemelmans et al., 2002). Primacy effects may be relevant to this paper's argument because the initial items in the test from a particular portion of the syllabus being tested will be retained longer in the STS buffer aiding recall. Recency effects may be relevant to this paper's argument because the section of the syllabus relating to the last or present item being attempted by the candidate being freshest in the STS may be recalled more readily for subsequent items. However, if the subsequent item comes from a different section of the syllabus, the STS buffer must now be replaced with newer content relevant to the new item. Recency as an effect is now attenuated and now less likely to occur. Hence an increased cognitive workload occurs for each candidate due to this perpetual random transmission between sections of the syllabus in computer-based testing contexts. This may be expected to hinder test-takers in the computer-based testing modality.

Support for the relevance of the attenuation of the recency effect to this paper may be found in a paper by Talmi and Goshen-Gottstein (2006). These authors cite an earlier paper in which they found that randomised test items in a multiple-probe recognition procedure will result in the attenuation and possible elimination of the recency effect. **Multiple-probe** procedures test for items to be recognised from a list by randomly presenting those items among distractor items and items from earlier lists which are also distractors for the present list. The **single-probe** approach differs from the multiple probe in that a study list of items to be recognised is followed by a test containing only items to be recognised and distractor items that did not feature in earlier recognition lists (Talmi & Goshen-Gottstein, 2006). Although the contexts are different and in the cited study "items" are entities on a list to be recalled not multiple-choice questions, I believe the multiple-probe recognition procedure is synonymous with the many different sections of the test syllabus. Thus the

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

randomised sequence in which test items draw from these different sections attenuates the recency effect impeding effective and efficient recall for the test-taker. The suggestion to use a single-probe approach (Talmi & Goshen-Gottstein, 2006) is reported to avoid the "test order confound". In the context of computer-based testing I infer that a single-probe approach can be obtained by ensuring multiple-choice items are grouped according to the section of the syllabus from which they are sourced. The logical question to ask is can grouping multiple-choice items from a section of the test syllabus together avoid a "test order confound" by modelling a single-probe approach for each section being assessed?

Memory and recall are key faculties a test-taker must rely on in testing contexts. Primacy and recency effects explain an apparent increased likelihood for items to be recalled. Of concern is the possibility that randomising test items may cause attenuation of the primacy and recency effects, diminishing accurate recall. In high stakes testing scenarios the effects will be more pronounced (Hancock, 2001; Keogh & French, 2001). Nevertheless these are the situations in which fairness and an ability to accurately and efficiently recall study material is even more critical due to the impact of the test performances on the futures of individual test-takers.

To sum up the overall affect for test-takers is the difficulty to practise good test taking habits that may affect them emotionally (Mealey & Host, 1992; Schutz et al., 2002). Coupled with this the sequential and coordinative task complexities introduced may affect their cognitive functionality (Chapell et al., 2005), and both of these issues are a direct result of the randomisation of test items in computer-based test assessments. Furthermore cognitive capacity may be affected by a weakening of the primacy/recency effects.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

3.5. Method

The paper is based on a previous study investigating the effect of item sequence randomisation on the performance of test-takers (Marks et al., in press). That study investigated the Item Randomisation Effect due to the sequence of difficulty of items randomly presented to students. The graphic representations shown in that paper resulted in reflections and in the subsequent effects forming the bulk of this chapter. It should be noted that these effects have not as yet been investigated but the commentary on them in this chapter will hopefully motivate future research focused on quantifying the actual affect on test-takers in computer-based test contexts.

Originally, existing computer-based test data was used to perform a pilot study. The data consisted of four sets of tests for 103 veterinary students in a year long course (Marks et al., in press). The first two tests were quite small consisting of twenty items and twenty-five items respectively. The last two tests were larger consisting of forty-nine and forty-eight items. It is from graphic analysis of the last two tests that one becomes aware of the trend of easy and difficult items following on from one another. The initial study of Marks et al. (in press) attempted to investigate the correlation between test-taker performance and the sequence of item difficulty presented to each student. To do this, normal performance for each candidate was defined as the average for that candidate for the four tests. The average difficulty index for each test was the average score for the test by the group. Item difficulty was defined as the ratio of the number of correct responses to the total number of attempts for the item (Reise & Henson, 2003). This implies that the closer the index is to unity the easier the item. A difficulty index of unity implies all candidates attempting that item correctly chose the key, and none selected any of the item's distractors. Conversely a zero score for the difficulty index is obtained if all of the attempts were incorrect, possibly an item that is very difficult or needs to be edited because it is ambiguous or flawed in some manner. Any candidate that scored significantly higher/lower

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

(deviations of at least fifteen percent considered significant) than their normal performance in any of the four tests was flagged. The sequence of items was plotted against the item difficulty index for these flagged performances. These graphs, with respect to the manner in which the candidates are presented easy and difficult items in a random fashion, will be presented and discussed later in this chapter.

The four tests are labelled Test 1, Test 2, Test 3, and Test 5. The fourth test was abandoned after an hour due to computer glitches at the venue. A totally new set of test items was selected and presented to the students the next day and was therefore labelled Test 5, but only four sets of test data were useful for the purposes of this study.

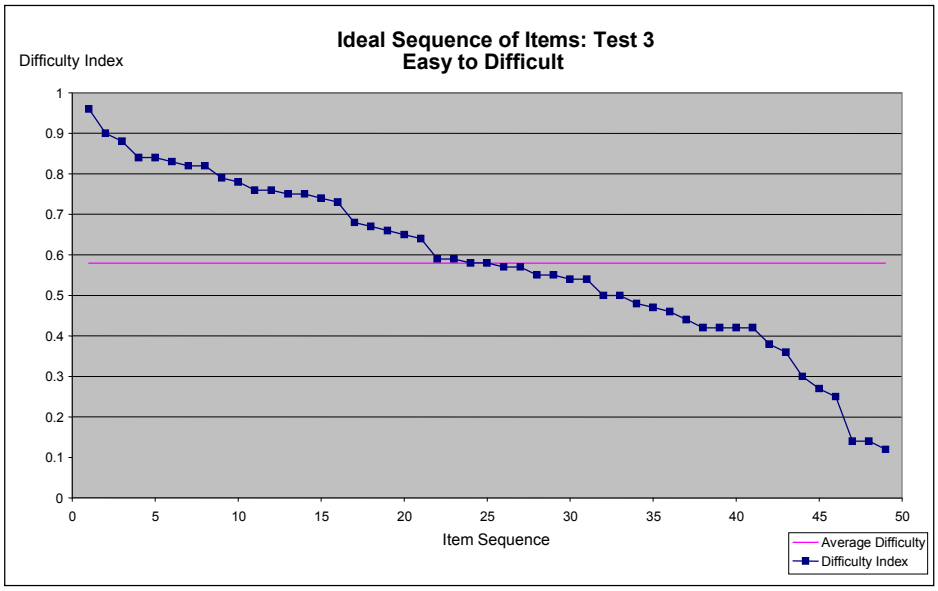

Figure 8: Ideal sequencing of test items using Test 3 data

An issue of interest for this paper is how often the graphic plot of item difficulty crosses over the average difficulty index for each test-taker. In an ideal computer-based test, (see Figure 8 above), the average difficulty line would be crossed only once about half way through the sequence of items presented to each candidate, and this would be with the initial items consisting of the easier items, and the latter items consisting of the more difficult items. Conversely the more often the average difficulty line is

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

crossed; the more problematic it may be for test-takers as this deviates most from what may be considered as good test-taking habits.

3.6. Discussion from Findings

The figures that follow are the most striking examples from each of the tests of the trends that become apparent, when one studies the data pertaining to the candidates that showed a significant difference in performance from their normal performance for the module across the four tests. Before commenting specifically on each figure, some common points of interest should be noted:

1. Diamonds indicate the Difficulty Index per Item,
2. The horizontal line labelled Average Difficulty is also the average scored for the particular test,
3. A line connects the diamonds such that the items are connected in the sequence presented to the test-taker,
4. The line connecting the diamonds can now conveniently show how often the average difficulty line is traversed,
5. The candidate's deviation from normal is indicated under the chart title.

Only three candidates from Test 3 deviated by fifteen percent or more from their normal performance, and only two for Test 5, so all these graphs will be reproduced and briefly commented on.

For Figure 9 (next page) the average difficulty line at 0.58 was crossed about twenty-one times in the test. The test contained only forty-nine items. The ideal of crossing this line only once as depicted earlier in Figure 8 is far exceeded by crossing more than twenty times. This candidate scored fifteen percent below their normal score for the four tests. This continual changing of difficulty level for successive items may be an emotional roller coaster ride throughout the test for this test-taker.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

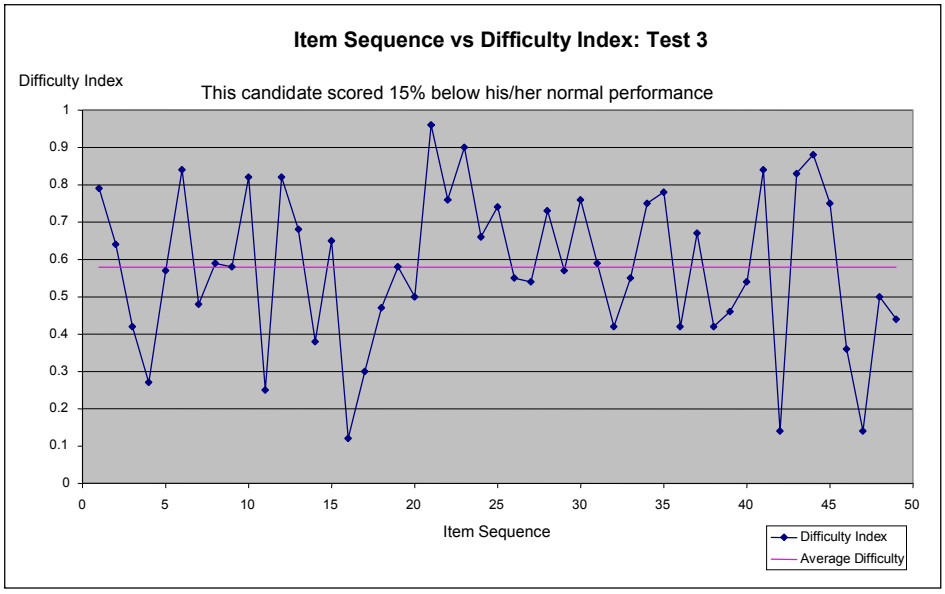

Figure 9: 1st Example from Test 3

In Figure 10 below the average difficulty line was also crossed about twenty-one times yet the candidate did significantly better than his/her normal performance. This may seem to counter the arguments forwarded thus far, however, the candidates scores over the four tests is worthy of comment.

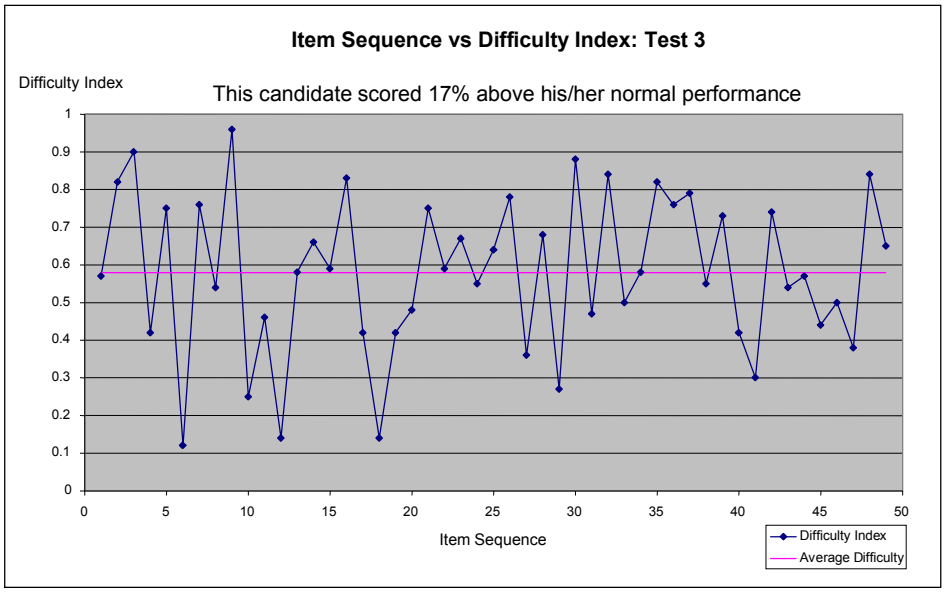

Figure 10: 2nd Example from Test 3

The candidate obtained 39% for Test 1, 29% for Test 2, 65% for Test 3, and 58% for Test 5, averaging only 48% for the four tests. He/she obviously

created using BCL easyPDF Printer Driver
Click here to purchase a license to remove this image

prepared more thoroughly for the last two tests in an attempt to convert a certain fail grade into a passing grade, hence the improved scores in Tests 3 and 5. Perhaps this candidate was disadvantaged by the emotional roller coaster effect of constant changes in difficulty index, and in fact might have done even better in this test if a more ideal pattern of item difficulty had been presented (see Figure 8).

For Figure 11, the average difficulty index line was crossed about twenty-five times. The candidate scored sixteen percent below his/her normal performance for the four tests. This roller coaster effect may be part of the cause for this deviation as the contrast with Figure 8 is obvious.

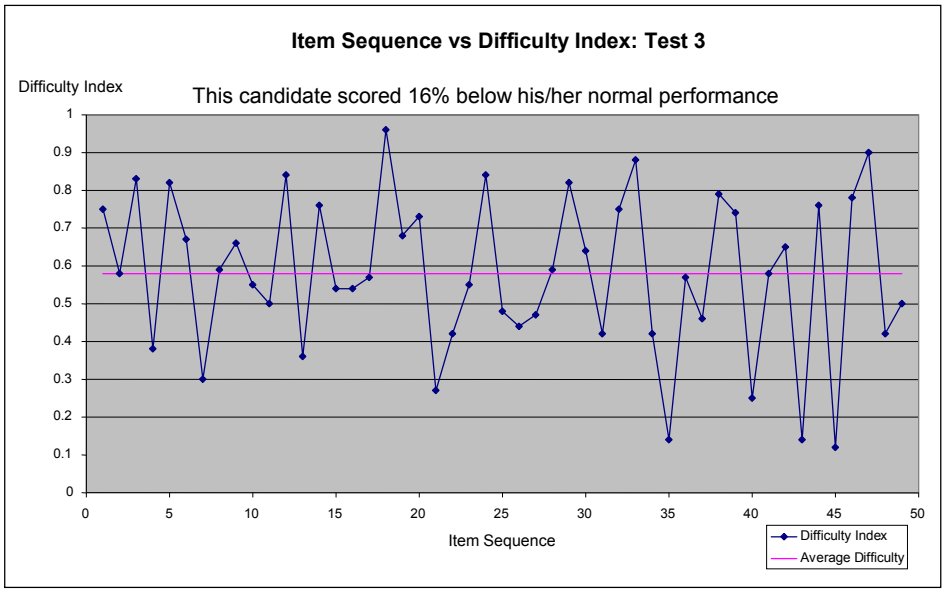

Figure 11: 3rd Example from Test 3

In Figure 12 (next page) one can see that for Test 5 the average difficulty index is also 0.58, however, it consisted of forty-eight items. The candidate scored fifteen percent below his/her normal performance. The average difficulty line was crossed about nineteen times, far in excess of once halfway through the test, the ideal, as shown in Figure 8.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

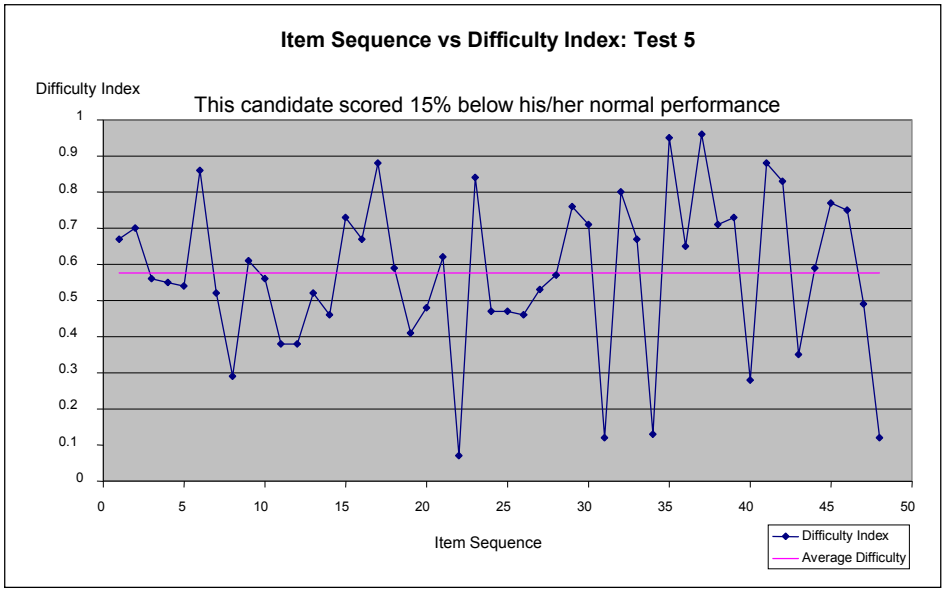

Figure 12: 1st Example from Test 5

In Figure 13 below, the candidate scored seventeen percent below his/her normal level, and the average difficulty line was crossed about twenty-four times. This is far from ideal, and it is my opinion that this could be affecting the candidate adversely. This criss-crossing pattern exists for all candidates for all the tests, so it is important to note that it isn't the only reason for the deviation from normal, but it certainly may be a contributory factor, that is particularly likely to affect the test anxious test-taker more severely.

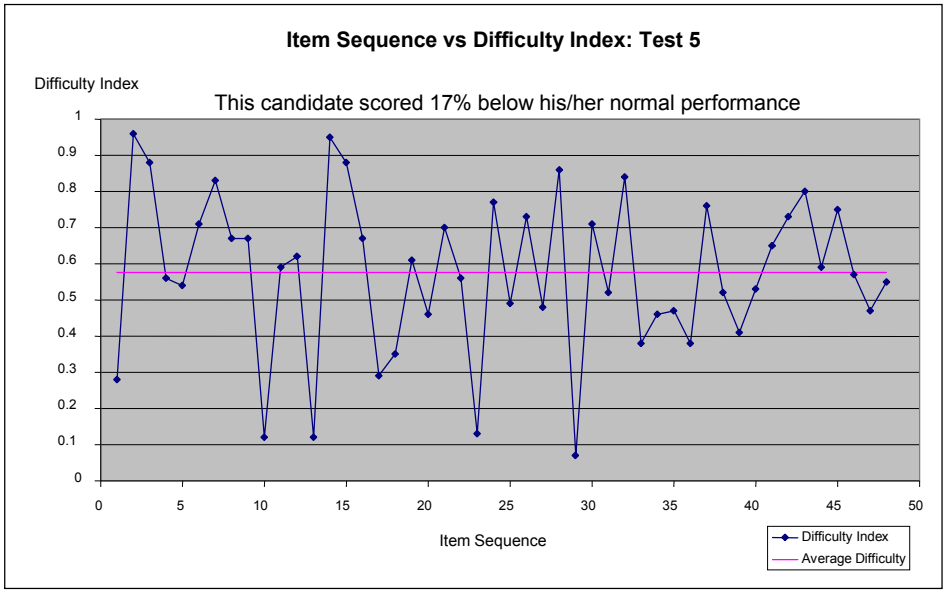

Figure 13: 2nd Example from Test 5

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

An additional factor mentioned earlier in the chapter, but not used as part of the discussion thus far is worthy of mention at this stage. The items are randomly presented to each candidate such that items from logical groupings of the syllabus are randomly dispersed throughout the test. By way of illustration, assume a test syllabus for a subject called Art 111, is made up of four sections. A total of 50 test items are to be presented to the test-takers. Assume twenty items are quizzing knowledge from section A of the Art 111 syllabus; the other three sections of the syllabus each have ten items, thus making up the 50 items of the test from sections B, C, and D respectively. A sensible test would ensure that all twenty items testing section A of the Art 111 syllabus should be presented consecutively one after the other. Section B has ten items that should also be presented to the test-taker consecutively, followed by the ten items from section C, followed by the remaining items from section D. A less sensible way to administer the Art 111 test items may be to maintain the sequence of items within the sections as described but to randomise the order in which the sections are presented, i.e. a section order of B-D-A-C. I now argue that the least sensible administration of the 50 items would be to ignore the four sections of the syllabus and randomly present the 50 items so that the four sections are quizzed in an illogical, erratic sequence thus: A-C-B-C-D-A-B-A-D-C-A-B….etc. Please note that present randomisation of test items in computer-based tests would use the sequence of items referred to as “least sensible” in the illustration above, yet test-users seemingly have not noticed this or are ignoring it. This paper wishes to ensure that those that have not noticed are made aware of the Item Randomisation Effect, and that those ignoring it should rather be calling for further research to ascertain whether it is acceptable to continue ignoring it. The future research should be focused on investigating the extent to which (i) cognitive workload, and (ii) memory and recall are affected.

Cognitive workloads are a function of either sequential complexity or coordinative complexity (Dutke & Stöber, 2001). The randomisation of items with no regard for the sections of the syllabus would certainly cause a sequential complexity to exist in that working memory is updated with independent information required for the subsequent item from a different

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

section of the syllabus. Coordinative complexity is also introduced as the processed results of previous items may be useful in attempting later items. However, as working memory is required to process data from various other sections of the syllabus the processed data from the earlier item may have already been purged from the working memory due to the sequential complexity introduced through randomisation of the items in a "least sensible" manner. The ideal would be for these items to be grouped together such that the working memory can be purged once all items for a section of the syllabus have been attempted. Due to the randomisation of test items it is quite likely that items from the same section are widely dispersed throughout the test resulting in both sequential and coordinative complexities taxing the working memory of test-takers throughout the assessment. In paper-based tests the items would be logically grouped according to their sections in the syllabus, in this way less of a coordinative complexity exists due to the candidate knowing that the working memory can be purged and updated anew with different information when items from a different section are attempted.

Memory and recall may be affected by randomisation of items through the attenuation of recency and primacy effects (Talmi & Goshen-Gottstein, 2006). The relevance of **recency** and **primacy** has been described in some detail in the literature review section of this chapter (see pages 66-68). This chapter merely serves to highlight that there are various concepts that may be relevant to computer-based testing practice, and that further research is needed to determine the need for fine tuning our randomising algorithms used in administering tests to our test-takers.

3.7. Recommendations

3.7.1. To Randomise…

The random item presentation algorithm must be programmed with far more care and consideration of the effects presented in this chapter. My

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

recommendation is for the test items to be sorted into three groups according to difficulty using each item's difficulty index. The three groups then consist of the easiest items, the items of average difficulty, and the toughest items. Each test-taker is presented with the test items such that all the easiest items are administered before any of the average difficulty items, and all the average difficulty items are presented before the set of toughest items. The items in each grouping are now randomised to ensure test security but the three groups of items are administered to test-takers such that the test becomes progressively more difficult. This option is easily implemented by test developers, and would certainly negate the potential adverse affects of huge fluctuations in difficulty index from one item to the next whilst maintaining a degree of randomisation as a means to counter cheating in computer-based test venues.

However, ensuring specific sections of the syllabus are kept together, whilst simultaneously randomising the easy, average, and difficult items as described above, is not easily attained.

3.7.2. Or Not to Randomise?

The obvious cause of the concerns highlighted in this chapter, is the randomisation of multiple-choice question items during computer-based tests. The recommendations that follow are ideas that may be used to eliminate the need for items to be randomised. In other words, cheating in the computer test venues needs to be eliminated.

Test venues should be radically kitted out or modified to ensure cheating is not possible. This would likely require a form of partitioning screen to prevent test-takers viewing neighbouring computer screens. However, this partitioning would shield candidates from invigilators making it easier for candidates to use illegal aids such as formulae sheets, or class notes. In addition this would be quite an expensive solution. Although it is not a likely solution, it is mentioned here as some institutions may prefer such an option.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

Another possible solution is to ensure neighbouring test-takers are attempting tests from different subjects' e.g. engineering faculty candidates are seated next to medical faculty candidates, and this sequence is alternated throughout the venue. This would cause other problems in terms of test timetables, and better planning and organisation by the team of computer technicians that set up computer venues in readiness for testing. However, this too may be an attractive option to some test users.

Variations of these recommendations could be used, with the aim of minimising the potential for cheating to occur in computer-based test venues, so that random presentation of items is no longer necessary.

3.8. Conclusions

The roller coaster analogy is useful to picture the potential effect on candidates with regard to cognitive workload, a lowered capacity to recall, and possible fluctuations emotionally.

It is hoped that this chapter will inspire researchers to create studies specifically suited to obtain correlations that will confirm or refute the item randomisation effects introduced herein.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

4. Chapter 4: Conclusion

This chapter concludes the pilot study into what is called Item Randomisation Effect, a term coined during this research as a result of the identification of a gap in the existing literature. The study is now briefly summarised, followed by discussion in the form of reflections. Lastly recommendations will be made for future research into this effect, possible amendments to be considered for present policy, and suggestions for improved assessment practice.

4.1 Summary

When I first started talking to peers about this particular study and what I wanted to research, I received encouragement from most, but from a few I got a surprising response. Some listened to the proposed study and told me that the answers to my research were already obvious to them and that this may not be a worthwhile study. Others implied that this research was not worthwhile because if it did actually show some test-takers to be disadvantaged, they were so few as to not warrant any corrective actions. Hence, my main question was posed to frame and even validate the need for this study. The main question: **Is it morally and ethically acceptable to ignore factors causing increased test anxiety if it only affects a small number of test-takers in computer-based assessment, or is one adversely affected student one too many?** The literature review emphatically answers this question. It is not acceptable to ignore factors that may cause even one test-taker to be affected by the test or by the manner in which it is administered. Hence the study was validated and other questions could be posed for the study.

Other questions asked, does present computer-based testing practice actually affect student performance, and if so can the practice be moderated to attenuate the affects and can the affects be attributed to either increased test anxiety or increased cognitive workload? Chapters 2 and 3 report on these

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

issues in some detail. Briefly, instances were found in Chapter 2 where it may be inferred that individual students have been affected by the difficulty sequence of the items presented to them in one of the four computer-based tests. The discussions that followed, particularly, in Chapter 3 indicate that the effects may be attributable to Item Randomisation Effect, but that more research would be required to substantiate these claims.

4.2. Discussion

4.2.1. Methodological Reflection

In the introductory chapter the following null hypothesis was stated:

"There is no significant difference between the performances of students presented difficult questions at the start of a randomised multiple-choice question computer-based test, compared to those presented the easy questions first."

No significant correlations were found during statistical analysis by the Statistics Department; hence the null hypothesis is NOT rejected. However, the ideographic nature of this study is the reason why statistical analysis is not likely to find any significant difference even if they may exist. Further research is therefore needed before one can confidently accept/reject the stated null hypothesis.

Originally my proposal called for an experimental research design that would consist of experimental and control groups to be tested such that statistical analysis could be performed and the results thus reported and discussed. However, this idea was considered to be worthy of a doctoral study, and was thus too broad for the purpose of this study. After some discussion it was decided to limit the scope of the study to that of an exploratory pilot study using existing computer-based test data. The data would be analysed to

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

determine if further research may be warranted. Hence an ex post-facto design was adopted.

Within this context, the study perhaps suffered from the decision to use existing data. Firstly, the use of existing data meant that statistical correlations were unlikely to be obtained. Statistical indices are suited to finding correlations that exist for the majority of a sample. This study wanted to identify exceptions that may only occur by chance. This is referred to as *Error Variance* in literature and was adequately discussed in Chapters 1 and 2. As a result this is an ideographic study and it cannot be anything else. An experimental design, could perhaps allow the study to be nomothetic with an ideographic concern. An experimental group could be administered a test such that the most difficult items are made to occur in the initial stages of the test, so the easiest items would be administered last. All these test-takers would then be interviewed to ascertain their feelings concerning the test. The results for each of these test-takers could be statistically analysed to see if statistically significant deviations can be obtained from the results of the control group. The control group should receive the items in a more ideal sequence, easy progressively getting tougher during the test. Interviews could also be conducted with this group for a qualitative slant to enhance possible statistical correlations. Variations on this design could be to swap the control and experimental groups in the sample so that each individual in the study experiences a test with an ideal and non-ideal sequencing of the test items. The statistical and qualitative analysis of the data after this swap may prove to be very useful. The study would be nomothetic because all persons in the sample would be experiencing similar item difficulty sequencing be it ideal or non-ideal. The ideographic concern remains because the chances of obtaining ideal or non-ideal item difficulty sequences in an actual assessment are quite remote. However, the ideal of ensuring not even one test-taker is affected, requires the results of such an experimental study to motivate for present item randomisation practice to be modified.

Lastly, the sample group of test-takers used for this pilot study was a group of senior veterinary undergraduate students. Veterinary students may be seen

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

as gifted in comparison to other groups of students at tertiary institutions. Veterinary students are likely to be mature test-takers with excellent test-taking skills. Certainly they are not likely to be prone to suffer from test anxiety. The Item Randomisation Effect is expected to be attenuated by test-takers with good test-taking habits, and so the decision to use such students as the sample for this pilot study was a poor one. Hindsight is twenty-twenty they say, but future researchers would do well to choose their sample with more care than I did.

4.2.2. Substantive Reflection

Up until this study, the effect of item randomisation on test-takers has been investigated but not with the angle considered in this study. This study is ideographic. Other studies sought to show equivalence across the modalities of computer-based and paper-based tests from a nomothetic stance. The ideographic perspective taken in this study is concerned with the performance of individuals taking the test, and how their performance is impacted through the Item Randomisation Effect. The nomothetic perspective is concerned with the performance of the test as administered in the computer-based modality, to ensure it will be seen as equivalent to the same test when administered in a different test mode.

The results of previous research are not relevant to this study as this study has a very different purpose. No studies could be found that investigated the effect of item randomisation on the performance of individual test-takers from the perspective of considering the sequence of item difficulty.

4.2.3. Scientific Reflection

This study has looked at the practice of randomising items in a computer-based test with respect to the difficulty level of each item and the sequencing of items throughout the test. The perspective of this paper is novel and to my knowledge has not been done by other researchers. The findings from

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

Chapter 2 are important and could be instrumental in motivating future research into the Item Randomisation Effect. The perceived gap in the existing literature has prompted me to adopt the term Item Randomisation Effect and I hope to see it used in future literature on this topic. The findings reported in Chapter 2 may be explained in terms of increased levels of test anxiety with an associated lowering in cognitive functioning. The ideographic perspective of this paper is important as it is what has allowed this gap in the existing literature to be noticed. Now it will be possible for future research to zero in on the potential for hindering individual test-takers from obtaining the test result they deserve. If testing practice can be improved then this pilot study has made a significant contribution to education and training.

The third chapter reports on further possible factors that can contribute to explaining why Item Randomisation Effect could be affecting individual test-takers. This chapter is the result of further reflection and reading upon depicting the results of Chapter 2 in graphic format. In Chapter 3 the contribution to the scientific body of knowledge could be in the way the effect is now also linked to cognitive psychology through the increase of cognitive workload, and through the attenuation of what is called **primacy** and **recency**, terms relating to memory and recall.

At this time it is important to note that the effects reported in this study are probably not limited to computer-based test modes. It is likely that the reasons why Item Randomisation Effect may be affecting individuals in the computer-based modality could be as valid for explaining similar affects to test-takers in other testing modalities. Therefore the contribution to the literature is broader than just effects prevalent in computer-based testing, and could be relevant to assessment in general.

4.3. Recommendations

The literature and existing research indicates that equivalence exists across the modalities that are acceptable and fall within the guidelines set out by

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

groups such as the American Psychological Association. However, the sequencing of questions in other test modes should also be investigated with respect to the difficulty levels of each question and its relative position or sequence it is administered to a test-taker. My opinion is that test users need to ensure that they are cognitive of the possible Item Randomisation Effect and how it could be relevant in other testing modes so that they can ensure test-takers are not hindered in scoring the result each one deserves. Perhaps the term could be expanded to include other test modes and question types, e.g. Question Difficulty Sequence Effect. My hope is that educators constructing tests will in future share the responsibility together with the test-taker to ensure that each test experience is able to obtain a fair, valid, reliable score representing the abilities of each and every test-taker.

There is now no excuse for researchers to continue to ignore the Item Randomisation Effect. It must be researched through properly designed research studies and the results thereof can then be used to show that either no action or change is warranted or to motivate adjustments to present computer-based testing practices are made.

The outcomes of future research could then be instrumental in shaping educational policy and practices of the future. Certainly, the recommendation coming out of this study is that test developers and test users must take care to ensure the randomising algorithm in future testing software is programmed to limit the chances of the Item Randomisation Effect occurring in future assessments.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

5. Acknowledgements

Thank you to the University of Pretoria, Computer Testing Section, for their assistance and for supplying the test data.

Thank you to the Faculty of Veterinary Sciences for permission to use their data.

Thank you to the Department of Statistics for the formal data analysis, particularly Mrs. Jaqui Sommerville.

Professor Cronje, thank you for your encouragement and support. Thank you for calling me to enquire my progress after I became dormant for extended periods. Several other supervisors would not do this, and then this thesis would not be contributing to the research community as I expect it will.

Thank you especially to my spouse and children for their understanding and patience. Thank you for making it easier for me to study by going away to visit distant family for several weekends without me.

Lastly, when I had no energy, when circumstances continued to conspire against me, when I had given up hope, I could always turn to Jesus and pray for guidance and strength and somehow each time I was empowered to overcome the inertia required to start again, and make significant progress. Thank you, Jesus, for your love and strength and presence throughout.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

6. References

Use parent nights to improve student test-taking skills. (2004). *Curriculum Review, 43*(5), 6.

Reduce Test Anxiety to Improve Student Performance. (2005). *Teaching Professor, 19*(9), 5.

Alessi, S. M., & Trollip, S. R. (2001). *Multimedia for Learning: Methods and Development* (3rd ed.). Boston: Allyn & Bacon.

APA. (1988). Rights and Responsibilities of Test Takers: Guidelines and Expectations.

APA. (2004). Code of Fair Testing Practices in Education.

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience, 4*(10), 829.

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.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

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.

Chapell, M. S., Blanding, Z. B., Takahashi, M., Silverstein, M. E., Newman, B., Gubi, A., et al. (2005). Test Anxiety and Academic Performance in Undergraduate and Graduate Students. *Journal of Educational Psychology, 97*(2), 268.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology, 33*(5), 593.

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), p20, 24p.

Gibson, C., Hoole, G., & Passchier, B. (1989). *Pass Your Exams Easily*.Cape Town: Struik.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

Glenn, R. E. (2004). Teach Kids Test-Taking Tactics. *Education Digest, 70*(2), p61, 63p.

Hall, M. E. (2000). A streamlined future for assessment. *Thrust for Educational Leadership, 29*(5), p15, 11p.

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.

Keogh, E., & French, C. C. (2001). Test anxiety, evaluative stress, and susceptibility to distraction from threat. *European Journal of Personality, 15*(2), 123.

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.

Marks, A. M., Cronje, J. C., & Mostert, E. (in press). Randomised Items in Computer-based Tests: Russian Roulette in Assessment?

Mealey, D. L., & Host, T. R. (1992). Coping with test anxiety. *College Teaching, 40*(4), 147.

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.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

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.

Ricketts, C., & Wilks, S. J. (2002). Improving Student Performance Through Computer-based Assessment: insights from recent research. *Assessment & Evaluation in Higher Education, 27*(5), 475.

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.

Russell, M., & Plati, T. (2002). Does it Matter With What I Write? Comparing Performance On Paper, Computer and Portable Writing Devices. *Current Issues in Education, 5*(4), 24.

SAQA. (2007). Design and develop outcomes-based assessments. In D. o. Education (Ed.).

Schutz, P. A., & Davis, H. A. (2000). Emotions and Self-Regulation During Test Taking. *Educational Psychologist, 35*(4), 243.

Schutz, P. A., Davis, H. A., & Schwanenflugel, P. J. (2002). Organization of Concepts Relevant to Emotions and Their Regulation During Test Taking. *Journal of Experimental Education, 70*(4), 316.

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), p292, 295p.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

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.

Varughese, J. A. (2005). TESTING, TESTING. *University Business, 8*(4), 59.

Wang, T. H., Wang, K. H., Wang, W. L., Huang, S. C., & Chen, S. V. (2004). Original article Web-based Assessment and Test Analyses (WATA) system: development and evaluation. *Journal of Computer Assisted Learning, 20*(1), 59.

Zenisky, A. L., & Sireci, S. G. (2002). Technological Innovations in Large-Scale Assessment. *Applied Measurement in Education, 15*(4), p337, 326p.

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image

7. Appendices

7.1. Appendix 1

Letter of Informed Consent from Dr. Peter Irons of Onderstepoort

Section Reproduction
Dept. Production Animal Studies

Tel. (012) 529 8218
Fax: (012) 529 8314
August 30, 2005

Mr. Anthony Marks
12 Vaalharts Str.
Brackendowns, Alberton
1448

Mr. Marks

Permission to use CBT data for research

I have no hesitation in allowing you to use our data for the purposes of your project under the circumstances you describe. The confidentiality of the information and particularly of the questions themselves is however of the utmost importance to us.

It would be of interest to us to see the results of your research, and a very direct benefit to us if you could provide us with electronic versions of our data, as I assume you will need to prepare for the purposes of your study. A weakness of the current CBT software is the inability to easily capture the analysis of the facility, discrimination and frequency of selection of specific options each time a question out of a library is used. Secondly, questions put into a specific library for a test are usually placed in random order when one reopens the test, meaning that the question loses whatever unique identifier number may have been given to it, and the numbering and order may be different each time one works with it. Therefore, to analyse the facility and discrimination stats of a question, one has to manually go through the stats, make sure they pertain to the correct question, then capture them into a spreadsheet for comparison with previous occasions on which the question was used. We have not done this due to the time involved, but would be very appreciative if you could do this for us during the course of your project. This would make this into a classic ‘win-win’ situation. However, any contribution towards making our question database more organised, accessible and easy to work with would be of benefit to us.

Please do not hesitate to contact us if we can be of any further assistance.

Regards,

_____
Dr. Peter Irons BVSc MMedVet Dip ACT
Tel. (012) 529 8019
pirons@op.up.ac.za

created using
BCL easyPDF
Printer Driver
Click here to purchase a license to remove this image