International Journal of
Research and Current Development
Available Online at http://www.journalijrcd.com
vol. 1, Issue, 1, pp.11-16, November, 2016

ISSN: XXX-XXXX

RESEARCH ARTICLE

COMPARING EXAMINATION PERFORMANCE ACROSS THREE ONLINE ASSESSMENT FORMATS IN AUSTRALIAN MEDICAL STUDENTS

Monique Trigg[*1]., John James Barnard[2]., Peter Devitt[3] and Hannah Pham[4]

[1,2]Excel Psychological and Educational Consultancy (EPEC) Pty Ltd (www.epecat.com). P O Box 3147 Doncaster East, 3109, Australia

[3,4]Department of Surgery, the University of Adelaide, Adelaide, South Australia, 5005, Australia

ARTICLE INFO

Article History:

Received 15th September, 2016
Received in revised form 25th October, 2016
Accepted 23rd November, 2016
Published online 28th December, 2016

Keywords:

Assessment; Guessing; Measurement; Multiple-Choice; Short Answer Question; Option Probability Theory

ABSTRACT

The aim of this study was to compare the performance across three assessment formats in online test administration; Multiple Choice Questions (MCQs), Short Answer Questions (SAQs) and Option Probability Theory questions (OPTs) on a sample of 276 Bachelor of Medicine and Bachelor of Surgery (MBBS) students at the University of Adelaide, in order to evaluate the advantages and disadvantages of these formats. Students were divided into groups according to their year of study (Fourth or Fifth Year) and separated in sessions to counterbalance the modes of administration. All students completed parallel tests in each mode. Results showed that students performed best in the MCQ and worst in the OPT, regardless of the year of study. The OPT scores purport to provide information about individuals' guessing which cannot be captured through MCQs and hence through yielding a more precise estimate of cognitive ability, may be effectively utilised as a diagnostic testing tool.

INTRODUCTION

Testing arguably dates back to the Han Dynasty (206 BC–AD 220) in China, when a civil service examination, the Imperial Examination (IE), was implemented to select government servants and leaders (Niu, W. 2007). Similar practices spread to Western civilizations such as Britain by the early-nineteenth century, with Short Answer Questions (SAQs) and essays being the predominant form of testing used at this time (McArthur, 1983). Tests which implemented uniform administration and scoring procedures became known as standardized tests, and their use became common practice by the 1920s in order to gain an insight into the quality of teaching and learning in schools (Gallagher, 2003). As technology developed, so did testing, with multiple choice testing increasing in popularity by the mid-20th century. Multiple Choice Questions (MCQs) are now one of the most common forms of assessment utilised in education, placement, selection, certification, licensure and diagnosis (Ventouras *et al.*, 2010), and due to the perceived objectivity and efficiency they offer, MCQs have gradually been more commonly used in university testing and examination scenarios, across a myriad of disciplines including medical education (Hudson and Treagust, 2013). With an increased need for cheaper and speedier test delivery in the early-20th century coupled with astronomical developments in the accessibility of personal computers (PCs) as well as internet capabilities, computer-based testing yet again changed the landscape of testing (Thurlow *et al.*, 2010). As PCs increasingly made their way into education, the benefits of computer technology as a tool for Learning and teaching became evident in allowing for the development of critical thinking, problem solving, and assessing knowledge based on learning material (Ventouras *et al.*, 2010). Due to the sustained growth and popularity of computer-enhanced learning environment in education, many assessment programs have followed suite, utilising multimedia learning material to measure knowledge in various formats such as, in MCQs or SAQs via the use of PCs (Wainer and Thissen, 1993; Park and Choi, 2008).

Since the introduction of PCs in assessment, many testing programs use the MCQ format over any other type of assessment format to assess knowledge due to the ease of administration, objective scoring and dynamic reporting of results offered (Delen, 2015). The advantages of using MCQs have been widely described in the literature, where it has been demonstrated that MCQs can be used to assess higher order cognitive knowledge, understanding and skills (Haladyna, 2004; Ebel and Frisbie, 1991). However, some fierce criticism has been voiced against MCQs, with one of the most popular criticisms of MCQs is that a test taker can answer a question correctly through guessing (Le Roux, 1999; Barnett-Foster and Nagy 1996). The more traditional method of using SAQs as such still remains popular in many areas of assessment where detailed knowledge and thorough understanding may need to be assessed (Ventouras *et al.*, 2010). One school of thought has argued there may be more error in the reported scores derived from MCQ testing than SAQ testing due to the

*✉ ***Corresponding author:*** **Monique Trigg**
Excel Psychological and Educational Consultancy (EPEC) Pty Ltd (www.epecat.com). P O Box 3147, Doncaster East, 3109, AustraliaMaharashtra, India

possibility of guessing the correct answer in an MCQ (Rowley and Traub, 1977). For example, in a four-choice item, there is a 25% chance of randomly guessing the correct answer when the test taker has no knowledge of the topic. Thus, if a test taker responded correctly to 11 out of 20 items, it is reasonable to suggest they may have correctly guessed the answer some of the 11 items. It could be claimed that based on probability, for every three items answered incorrectly, one could be answered correctly through guessing, and therefore three of the 11 items answered correctly, were correctly guessed.

In general, a correction to the score can be calculated resulting in reduced scores (except for perfect scores). Herein lies a dilemma; a test taker would have a corrected score of zero if 25 of 100 four-choice items in a test were answered correctly. But if the test taker is a member of the cohort for who the test is intended, it is very unlikely that such a test taker will generally not know anything. For some items, the test taker may know the correct answer and in others they may have some partial knowledge. It is clear that implementing the correction for guessing formula does not adequately solve the issue of guessing in MCQs therefore. In this argument, it is assumed that the test taker has guessed randomly if the answer was unknown and that all incorrect responses are due to random guessing. This may be true for some items, but generally test takers are able to eliminate one or more options based on partial knowledge. This would result in increasing the chance of a correct guess which is now a more educated guess.

Rasch measurement philosophy purports to illuminate the issue of possible guessing by means of fit statistics (Andrich *et al*., 2012). In Rasch, a test taker's probability to correctly respond to an item is estimated by means of a monotonically increasing function, called the Item Response Function or IRF (Bond and Fox, 2007). Based on the concept of the score over all the items in the test being a sufficient statistic, two test takers who responded correctly to the same number of items will have the same ability estimate. However, one test taker may have guessed some items correctly while the other not or to a lesser extent.

Since person ability estimates and item difficulty estimates can be located on the same scale, the probability that each test taker has to correctly respond to each item can be calculated. It could therefore be suspected that a correct answer to an item with difficulty significantly above a test taker's ability estimate, was guessed correctly. This can be flagged by a fit index. Where most test takers have responded largely in accordance with the model's expectations, misfit of an individual test taker can be attributed to anomalous test taking behaviour of some kind. Whatever the underlying cause, a response vector which is inconsistent with an otherwise well-fitting model may indicate that the test, though possibly functioning well for the group as a whole, has failed to provide an appropriate measure of the relevant ability for that particular test taker.

The three-parameter "logistic" Item Response Theory (IRT) model includes an item parameter which gives an indication of the probability for any test taker to guess the correct answer (Hambleton and Swaminathan, 1985). The pseudo-chance (guessing) parameter estimates are typically less than the random guessing value (Hambleton *et al*., 1991). Some studies identified both theoretical and practical issues regarding the guessing parameter and estimation algorithms initially posed problems resulting in calibrations failing to converge (San Martin *et al*., 2006). Han (2012) noted, that many computer programs provide approximate average parameter estimates for the parameter in the case of non-convergence and actually mislead practitioners to interpret the outcome of the calibration as successful calibration.

The fact is that it will never really be known whether a test taker guessed the correct answer or knew the correct answer if an item was answered correctly unless the test taker is directly asked. In some cases the test taker would be quite sure about their answers whilst in other cases only to some extent or not at all. It can be assumed in most cases that the test taker evaluates the different options and if a single option is to be selected, the test taker selects the option considered to be correct or has the highest chance of being correct, according to the test taker's evaluation. If the test taker can eliminate one or more options, the chance of responding correctly increases and if the test taker does not know the answer at all, would most likely randomly guess if an answer is given.

It should be noted that it is not an item that guesses but rather a person, and therefore a guessing parameter needs to be a person parameter and not an item parameter or at least an interaction parameter. Several approaches have tried to capture the amount of guessing by test takers through scoring rules (De Finetti, 1970). For example, in confidence weighting the certainty of an answer is required from the test taker about their answer. This secondary scale has some value if the test taker chose the correct answer, but is otherwise rather useless. In probability theory the test taker assigns probabilities to each option of a MCQ to indicate the certainty that each option is correct. Although this overcomes the issue in confidence weighting, the main problem is that the score is influenced by how the probabilities are distributed. It can, for example, be shown that a test taker who rules out some options can score lower than a test taker who doesn't. Recently a new paradigm, Option Probability Theory (OPT) has found its way into the literature (Barnard, 2015; Barnard, 2012). In this theory the test taker assigns a percentage to any number of given possible answers and the percentage assigned to the correct answer is used to calculate a scaled score Ψ together with a realism index ρ which indicates how realistic the test taker is about their knowledge.

Instead of scoring a question dichotomously (correct or incorrect) like common practice in MCQ scoring, Ψ is based on a logarithmic scoring function that rewards higher percentages assigned to correct answers and penalises higher percentages assigned to incorrect answers. An estimated score, ε, is also calculated from the percentages assigned and interprets higher percentages assigned to correct answers as correct responses and low percentages assigned to correct answers as incorrect responses. The estimated score ε thus dichotomises the responses to yield response vectors that mimic MCQ scoring.

ISSN: XXXX-XXXX

It is expected that ε will be greater than Ψ for most if not all test takers and that the difference may become greater as a function of the number of questions to which higher percentages are assigned to incorrect answers. The realism index ρ measures whether the test taker's confidence in their choices matches their level of knowledge. Because the index is a measure of confidence, it penalises or rewards confident choices much more greatly than the Scale Score Ψ. Further, because the index measures both confidence and knowledge, and is sensitive to fallacies it is influenced more by incorrect answers than correct ones.

Study

The study was designed to compare the performance of medical students across three assessment formats in online test administration; Multiple-Choice Questions (MCQs), Short-Answer Questions (SAQs) and Option Probability Theory Questions (OPTs).

Participants

The study was conducted at the University of Adelaide, South Australia, in September 2015. The sample consisted of 276 MBBS students (157 females and 119 males). Participation was voluntary, and students were informed about the aims and the nature of the research prior to registering their interest to participate online via a registration website. Students were divided into two groups, according to the year of study: group A=136 students (72 females and 64 males) from the Fourth Year, and group B=140 students (85 females and 55 males) from the Fifth Year.

Instruments

A unique username and password was generated for each participant in order to access the three modules of the formal practice examination. These were only provided to participants on the day of their examination, upon sign-in. Two modules of the practice examination (MCQ and OPT) were administered online via EPEC's cloud platform Prognoser (EPECPrognoser, 2013), and the third module of the practice examination (SAQ) was administered through an application on the Google cloud platform, Google Forms. Participants were advised they were required to complete all three modules in order to obtain their results.

The three modules were constructed in parallel to cover the same content areas and the tests were constructed to be pair wise as parallel as possible. For Fourth Year Students, the disciplines included; medicine, surgery, psychiatry, orthopaedics and general practice/medicine. For Fifth Year Students, the disciplines included; medicine, surgery, psychiatry, orthopaedics, general practice, obstetrics/ gynaecology, paediatrics and APIC (Anaesthetics, Pain Medicine and Intensive Care).

The structure of both the MCQ and the OPT modules consisted of 40, five-option items each, which had to be completed within 45 minutes. For each item, there was one correct answer only. The SAQ module consisted of 40 items, which had to be completed within 90 minutes. All responses were captured online and collated for scoring and analysis. The responses from the SAQ module were collated and distributed to a team of six independent markers, which were marked according to a marking guide.

The MCQ and OPT results were quantified through marks from 0 to 40, with the SAQ results for Fourth Year ranging between 0 and 91, and results for Fifth Year ranging between 0 and 90.

Procedure

Upon online registration between July 2015 and September 2015 through an application on the Google cloud platform, Google Forms, all participants were asked whether they would prefer to sit the practice examination in a morning (08:00 through 12:00) or afternoon session (13:00 through 17;00). Each participant was allocated to their first preference. Each participant was subsequently provided access to an online practice test which consisted of five OPT practice questions to be completed on Prognoser. This allowed participants to familiarise themselves with the cloud platform, as well as the OPT question format. Access to the practice test closed on 4 September 2015. It is noted that only 110 of the 276 students completed the practice test.

Fifth Year Students sat the practice examination on 12 September 2015, and Fourth Year Students sat the practice examination on 13 September 2015. All participants were required to arrive 15 minutes prior to their examination start time for a briefing session. Each examination session (morning and afternoon on 12 September; and morning and afternoon on 13 September) were further halved, with one group sitting the OPT and MCQ modules first, and the other group sitting the SAQ module first. Within each session, the MCQ and OPT modules were also counterbalanced. For all participants, the OPT and SAQ modules contained items in a fixed order and the MCQ module contained items in randomised order.

RESULTS

Four sets of scores were available for analysis, namely MCQ scores, SAQ scores, OPT scaled scores Ψ and OPT estimated scores ε. To simplify direct comparisons, the four sets of scores were all converted to a 10-point scale.

Fourth Year Students

Table I Descriptive Statistics for Fourth Year Students

	Min	Max	Mean	Std. Dev
SAQ	1.7	6.9	4.2	1.1
MCQ	3.0	7.8	5.1	1.0
OPT ε	1.3	6.1	3.8	0.9
OPT Ψ	0	3.4	0.5	0.8

Fourth Year Students generally performed better in the MCQ format, followed by the SAQ format. The mean OPT Ψ performance was significantly lower than the MCQ and SAQ performances. A one-way repeated measures analysis of variance (ANOVA) was used to explore and compare the differences in more detail. There was a significant effect

ISSN: XXXX-XXXX

for mode of administration, Wilks' Lambda = 0.038, F (3,133) =1112, p < 0.0005, multivariate eta squared = 0.962.

Table II Correlations amongst the Four Sets of Scores of the Fourth Year Students

		SAQ (10)	MCQ (10)	OPT tot (10)	OPT (10)
SAQ (10)	Pearson Correlation	1	.590**	.362**	.273**
	Sig. (2-tailed)		.000	.000	.001
	N	136	136	136	136
MCQ (10)	Pearson Correlation	.590**	1	.345**	.259**
	Sig. (2-tailed)	.000		.000	.002
	N	136	136	136	136
OPT tot (10)	Pearson Correlation	.362**	.345**	1	.587**
	Sig. (2-tailed)	.000	.000		.000
	N	136	136	136	136
OPT (10)	Pearson Correlation	.273**	.259**	.587**	1
	Sig. (2-tailed)	.001	.002	.000	
	N	136	136	136	136

***Correlation is significant at the 0.01 level (2-tailed).*
OPT tot = OPT Ψ and OPT = OPT ε

More than 34% of the variance in the Fourth Year Students' scores between SAQ and MCQ and also between Ψ and ε is explained by the assessment format. This is a respectable amount of variance when compared to research of this nature conducted.

Fifth Year Students

Table III Descriptive Statistics for Fifth Year Students

	Min	Max	Mean	Std. Dev
SAQ	2.2	7.3	5.0	1.0
MCQ	3.5	8.0	6.0	1.0
OPT ε	2.1	6.5	4.7	0.9
OPT Ψ	0	4.9	1.5	1.1

As in the case of the Fourth Year Students, it is noted that the Fifth Year Students generally performed better in the MCQ module, followed by the SAQ module. The mean OPT Ψ performance was significantly lower than the MCQ and SAQ performances. A one-way repeated measures analysis of variance (ANOVA) was used to explore and compare the differences in more detail. There was a significant effect for mode of assessment, Wilks' Lambda = 0.056, F (3,137) = 766, p < 0.0005, multivariate eta squared =0.944.

Table IV Correlations amongst the Four Sets of Scores of the Fifth Year Students

		SAQ (10)	MCQ (10)	OPT tot (10)	OPT (10)
SAQ (10)	Pearson Correlation	1	.539**	.322**	.312**
	Sig. (2-tailed)		.000	.000	.000
	N	140	140	140	140
MCQ (10)	Pearson Correlation	.539**	1	.361**	.295**
	Sig. (2-tailed)	.000		.000	.000
	N	140	140	140	140
OPT tot (10)	Pearson Correlation	.322**	.361**	1	.606**
	Sig. (2-tailed)	.000	.000		.000
	N	140	140	140	140
OPT (10)	Pearson Correlation	.312**	.295**	.606**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	140	140	140	140

***Correlation is significant at the 0.01 level (2-tailed).*
OPT tot = OPT Ψ and OPT = OPT ε

As in the case of the Fourth Year Students, the results indicate that around 30% of the variance in the Fifth Year Students' scores between SAQ and MCQ is explained by the assessment format whilst around 37% of the variance between Ψ and ε is explained by the assessment format.

DISCUSSION

The students found the tests relatively difficult with means of 42% and 50% for the SAQ component respectively for Fourth and Fifth Year; 51% and 60% for the MCQ; 38% and 47% for the OPT estimated score ε and 5% and 15% for the OPT scaled score Ψ. Both cohorts performed the best in general in the MCQ tests and the correlations between the MCQ scores and the SAQ scores were above 0.5 and interpreted as large.

It is interesting to note the relatively low correlations (0.345 and 0.361 for Fourth and Fifth Year respectively) between the MCQ scores and the OPT estimated scores ε if it is taken into account that the estimated score ε was calculated by giving a full mark (regardless of percentage assigned) to confident and correct responses, reduced marks for responses that were not confident but correct and zero marks for incorrect answers. A possible explanation is that there was more than average guessing in MCQ and OPT since the tests were rather difficult and responses were not negatively marked.

Taking the magnitude of the correlations between the MCQ scores and the OPT estimated scores ε as a baseline, the correlations of the OPT scaled score Ψ and the MCQ score of 0.295 and 0.259 for Fourth and Fifth Year respectively are informative in the sense that they are not unusually low.

Although these statistics illuminate some general trends, more specific information lies within the details. Since SAQs are constructed response questions rather than selected response, it can be assumed that guessing is minimised when compared to MCQ and OPT as suggested by Ventouras *et al*. (2010). In contrast, SAQs may be more subject to inflated (or underrated) scores due to marking severity. To explore the differences in scores further, a few extreme cases were investigated.

The Fourth Year Student who attained the lowest realism index score had the following results: SAQ = 27.2%; MCQ = 30.0%; OPT ε = 37.5% and OPT Ψ = 0%. This student responded with 100% assigned to an incorrect answer to 25 OPT questions and thus assigned 100% to the correct answers of 15 questions. The student answered 12 of the MCQ questions correctly. With 100% confidence assigned to incorrect answers of 62.5% of the OPT questions, it can be inferred that the student guessed the answers to most of the MCQ as well as the OPT questions. The OPT penalty reflect in the extremely high negative realism index. It is noted that the student's SAQ was less than both the MCQ and the OPT ε scores.

The Fifth Year Student who attained the lowest realism index score had the following results: SAQ = 43.0%; MCQ = 58.0%; OPT ε = 39.0% and OPT Ψ = 0%. This student responded with 100% assigned to an incorrect answer to 21

ISSN: XXXX-XXXX

OPT questions and to four further questions with less than 50% assigned to the correct answer. Some credit was given to these four questions in the ε score. The student assigned 100% to the correct answer of only four questions and assigned 70% or more to the correct answers of 11 questions. These results suggest that the MCQ score was inflated due to some guessing.

The Fourth Year Student who attained the highest realism index score had the following results: SAQ = 44.4%; MCQ = 60.0%; OPT ε = 52.5% and OPT Ψ = 17.3%. This student responded with 100% assigned to a correct answer to 13 OPT questions; two further questions with more than 85% assigned to the correct answer and to 10 more questions with 50% assigned to the correct answer. The student assigned 100% to the incorrect answer of 11 questions and assigned less than 30% to the correct answers of four questions. The student had 24 correct answers in MCQ. If the 13 questions answered correctly with 100% confidence is added to the two questions with at least 85% confidence, it can be concluded that the student was sure about the answers of 37.5% of the OPT questions. Assigning 50% to the correct answer of a further ten questions increased the OPT ε score to 52.5%. The OPT Ψ score accounted for the 11 questions to which the student assigned 100% to an incorrect answer, and to a lesser extent to the four questions to which low percentages were assigned to the correct answers.

The Fifth Year Student who attained the highest realism index score had the following results: SAQ = 66.0%; MCQ = 75.0%; OPT ε = 64.0% and OPT Ψ = 34%. This student responded with 100% assigned to a correct answer to 21 OPT questions and to five further questions with more than 80% assigned to the correct answer. The student assigned 100% to the incorrect answer of 11 questions and assigned less than 30% to the correct answers of three questions. These results suggest that the student may have guessed the correct answer to at least four MCQ questions. The student's MCQ score was 75% which means that 30 of the 40 questions were scored as correct. The OPT ε score, however, suggests that the student was confident about only 26 questions answered correctly. The OPT Ψ scoring gave credit to the 21 items answered correctly with 100% confidence and also to the five questions to which high percentages were assigned to the correct answers. However, the student assigned 100% to incorrect answers of 11 questions which reduced the OPT Ψ score significantly, indicating some serious misconceptions or lack of knowledge.

CONCLUSIONS

The OPT ε score can be considered as the closest approximation of the 'true' performance since it credits high confidence to correct answers whilst allowing some credit to lower confidence of correct answers. In both Fourth and Fifth Year, the mean OPT ε score is 13% less than the mean MCQ score. This can be interpreted as the MCQ score inflation due to guessing, since incorrect answers in the MCQs were not penalised. This conclusion is further supported by the high correspondence between the OPT ε scores and the SAQ scores. This study also found that both groups of students performed better overall in the MCQ when compared to the SAQ, corroborating the findings of Hudson and Treagust (2013).

The OPT Ψ was significantly less than the other scores, almost exclusively due to penalising for high percentages assigned to incorrect answers. This observation should be made in context of the difficulty of the test. For example, if a student sat an easier 40-item OPT test and answered 35 questions correctly with 100% confidence and the remaining five questions with 50% confidence, the OPT ε score would be 93.8% and the OPT Ψ score 91.0%. If the test was a MCQ test, the student would most likely have answered the 35 items correctly (87.5%) and increased this score depending on which way the other five items were answered. If the student assigned 0% in the OPT version to the five items not answered correctly with 100% confidence, the OPT ε score will be 87.5% and the OPT Ψ score 75.0%. The OPT ε score is exactly the same as it would have been for the MCQ version whilst the OPT Ψ score penalises for the five items.

The study indicated that all students were disfavoured by the OPT examination format when compared with the MCQ and SAQ formats, regardless of the year of study. As such the design and use of the OPT testing format may in fact be more appropriate as a diagnostic testing tool rather than an achievement testing tool since it is able to clearly identify misconceptions, strengths and weaknesses, supporting the findings of Barnard (2015).

References

Andrich D, Marais I and Humphry, S. 2012. Using a Theorem by Andersen and the Dichotomous Rasch Model to Assess the Presence of Random Guessing in Multiple Choice Items. *Journal of Educational and Behavioural Statistics*, 37(3): 417-442.

Barnard, J.J 2015. Option Probability Theory: A quest for better measures. American Review of Mathematics and Statistics, 3(1): 61-69.

Barnard, J.J 2012. A primer on measurement theory. Melbourne: Excel Psychological and Educational Consultancy.

Barnett-Foster, D and P. Nagy. 1996. Undergraduate Student Response Strategies to Test Questions of Varying Format. Higher Education 32(2): 177–198.

Bond, T.G. and Fox, C.M. 2007. Applying the Rasch Model. Second edition. London: Lawrence Erlbaum. Chamorro-Premuzic, T, Furnham, A, Dissou, G and Heaven, P. (2005). Personality and preference for academic assessment: A study with Australian University students. Learning and Individual Differences, 15: 247–256.

De Finetti, B. 1970. Logical foundations and measurement of subjective probability. Acta Psychologica, 34:129-145.

Delen, E. 2015. Enhancing a Computer-Based Testing Environment with Optimum Item Response Time. EURASIA *Journal of Mathematics, Science & Technology Education*, 11(6): 1457-1472.

ISSN: XXXX-XXXX

Ebel, R. L and Frisbie, D.A. 1991. Essentials of educational measurement. New Jersey: Prentice Hall. EPEC Prognoser [Computer software] 2013. http://www.Epecprognoser.com.

Furnham, A, Batey, M and Martin, N. 2011. How would you like to be evaluated? The correlates ofstudents' preferences for assessment methods. Personality and Individual Differences, 50: 259-263.

Gallagher, C. J. 2003. Reconciling a tradition of testing with a new learning paradigm. Educational Psychology Review, 15(1): 83-99.

Haladyna, T.M. 2004. Developing and validating multiple-choice test items. London: Lawrence Erlbaum. Hambleton, R.K., Swaminathan, H. and Rogers, H.J. 1991. Fundamentals of item response theory. Newbury Park, CA: Sage Publications.

Hambleton, R.K and Swaminathan, H. 1985. Item Response Theory: Principles and Applications. Boston: Kluwer-Nijhoff.

Han, K.T. 2012. Fixing the c parameter in the three-parameter logistic model. Practical Assessment, Research & Evaluation, 17(1): 1-24.

Hudson, R. D and Treagust, D. F. 2013. Which form of assessment provides the best information about student performance in chemistry examinations? Research in Science & Technological Education, 31(1): 49-65.

Le Roux, I. 1999. Computer-based testing in higher education. MEd dissertation. Pretoria: University of retoria. McArthur, D.L. 1983. Educational testing and measurement: A brief history. Los Angeles: University of California.

Niu, W. 2007. Western influences on Chinese educational testing, Comparative Education, 43(1): 71-91. Park, J and Choi, B. 2008. Higher Retention after a New Take-Home Computerised Test. British *Journal of Educational Technology*, 39(3): 538-547.

Rowley, G.L and Traub, R.E. 1977. Formula scoring, number-right scoring, and test taking strategy. *Journal of Educational Measurement*, 14(1): 15-22.

San Martin, E, del Pino, G and de Boeck, P. 2006. IRT models for ability-based guessing. Applied Psychological Measurement, 30(3): 183-203.

Thurlow, M., Lazarus, S. S, Albus, D and Hodgson, J. 2010. Computer-based testing: Practices and considerations (Synthesis Report 78). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Ventouras, E, Triantis, D, Tsiakas, P and Stergiopoulos, C. 2010. Comparison of Examination Methods Based on Multiple-Choice Questions and Constructed-Response Questions Using Personal Computers. Computers & Education, 54(2): 455-461.

Wainer, H and Thissen, D. 1993. Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction. Applied Measurement In Education, 6(2): 103-18.

How to cite this article:
Monique Trigg., John James Barnard., Peter Devitt and Hannah Pham. 2016, "Comparing examination performance across three online assessment formats in australian medical students", *International Journal of Research and Current Development*, 1(1): 11-16.