American Review of Mathematics and Statistics
June 2015, Vol. 3, No. 1, pp. 61-69
ISSN: 2374-2348 (Print), 2374-2356 (Online)

Published by American Research Institute for Policy Development
DOI: 10.15640/arms.v3n1a7
URL: http://dx.doi.org/10.15640/arms.v3n1a7

Option Probability Theory: A Quest for Better Measures

John J Barnard[1]

Abstract

Most test takers do sometimes guess when responding to multiple-choice questions. Different theories and models attempt to address this issue in different ways. It is argued that test takers seldom randomly guess; they rather eliminate one or more options based on partial knowledge. Option Probability Theory (OPT) deals with guessing as a personal interaction parameter through requiring test takers to indicate how sure they are about each option as being the correct answer. Through calculating a realism index an indication can be obtained about how realistic a test taker has assigned these probabilities and as such serves as an indicator of the amount of guessing and certainty of answers. A small study suggests that OPT results are comparable to traditional results, but with significantly more information about test takers' knowledge and understanding. Practitioners can use this additional information to identify misconceptions, where there is uncertainty about answers and where guessing can be suspected. Using the realism index, a diagnostic analysis of response patterns can be made and detailed feedback can be given to test takers.

Keywords: Multiple-choice, scoring, measurement, assessment, option probability theory

1. Introduction

Millions of tests are administered on a daily basis and many of these use multiple-choice questions (MCQs). Items, as MCQs are commonly referred to, are used in tests for placement, selection, certification, licensure, diagnosis and for many other purposes (Haladyna, 2004; Phelps, 2000) and proved to be useful with many advantages over open ended questions (Ebel & Frisbie, 1991). Methods for scoring MCQs have progressed and became more sophisticated especially over the past three decades. Although simple dichotomous scoring (correct or incorrect) is still popular, partial credit and other scoring regimes have gained ground (Masters, 1982).The aim of this article is to discuss an alternative way of scoring MCQs.

2. Classical Scoring of MCQs

MCQs are typically scored dichotomously. A statement is made or a question is asked and usually four or five options (possible answers) are given. Type A items in which only one option is correct and the other options (distracters) are incorrect are the most popular type of MCQ and therefore the focus in this article will be on this type of MCQ. If the test taker selects the correct answer (key) a score of one is awarded – otherwise a score of zero is awarded to the response. In Classical Test Theory (e.g. Crocker & Algina 1986) the test taker's performance on a test is calculated by adding the item scores. The reliability of the test, calculated as the Kuder-Richardson formula 20 index, is used to calculate the standard error of measurement (SEM) which is used to indicate how precise the score is (Barnard, 2012). Note that the same constant value for the SEM is used to determine a range in which the test taker's "true ability" is located, irrespective of the performance. Any test taker has a chance of selecting the correct answer purely by guessing. There is, for example, a 20% chance that a randomly selected option will be correct in a five-choice item. Item writers usually ensure that the distracters are plausible answers and this has an effect on random guessing.

[1]EPEC Pty Ltd. / University of Sydney. Email: John@EPECat.com

But, guessing can inflate scores (e.g. Rowley & Traub, 1977) and therefore a formula for correcting scores for guessing was developed (e.g. Frary, 1988). It can be reasoned that a test taker who responded correctly to 11 out of 20 four-choice items may have guessed some of the 11 items correctly. In theory, for every three four-choice items answered incorrectly, one would be guessed correctly and therefore three of the 11 items answered correctly can be considered as if they were correctly answered through guessing. In general a correction to the score can be calculated as follows:

$$C = R - \frac{W}{k-1} \quad (1)$$

where C = corrected score, R = number of items answered correctly, W = number of items answered incorrectly and k = number of choices in the items. The test taker who had 11 correct answers will have a corrected score of 8. 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 may have some partial knowledge in others. It is clear that implementing the correction for guessing formula doesn't solve the issue of guessing (and scoring) in MCQs. 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.

3. Scoring MCQs In Modern Test Theory

In Rasch measurement (e.g. Bond & Fox, 2007; Andrich, 1988; Wright, 1977) MCQs are scored through estimating the probability that a test taker with a certain ability has to correctly respond to an item with a certain difficulty by means of a monotonically increasing function, called the Item Response Function (IRF) with mathematical form

$$P\{x_{ni} = 1/\beta_n, \delta_i\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \quad (2)$$

which simply states that the probability for a certain outcome to be one, given a person *n* with ability β and an item *i* with difficulty δ can be calculated by substituting the values into the equation with *e* the base of the natural logarithm. For example, a test taker with ability 0.2 log its will have a 55% chance of answering an item with difficulty of zero logits correctly. 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 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 which is commonly based on residuals in a χ^2 statistic. 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.

One of the main advantages of the Rasch approach over the Classical approach is that a SEM is calculated for each individual test taker instead of applying the same SEM to all test takers. This results in more realistic intervals in which test takers' ability estimates are located – a test taker who guessed less than another test taker will have a higher precision measure of ability. Whereas the Rasch model is based on a constant value for item discrimination, the two-parameter item response theory (IRT) model includes a parameter in the equation to account for variability in item discrimination. The three-parameter "logistic" IRT model further extends the modelling through an item parameter which gives an indication of the probability for any test taker to guess the correct answer (e.g. Hambleton, Swaminathan & Rogers, 1991; Hambleton & Swaminathan, 1985).

The basic mathematical form of the IRF accounts for this through addition of parameters (see equation 2):

$$P\{x_{ni}=1/\beta_n,\delta_i\}=\gamma_i+(1-\gamma_i)\frac{e^{D\alpha_i(\beta_n-\delta_i)}}{1+e^{D\alpha_i(\beta_n-\delta_i)}} \quad (3)$$

where α is the discrimination parameter, γ is the guessing (pseudo chance) parameter and D is a scaling constant. The γ-parameter estimates are typically less than the random guessing value. Some studies identified both theoretical and practical issues regarding γ and estimation algorithms initially posed problems resulting in calibrations failing to converge – see, for example, San Martin, del Pinoand de Boeck (2006).Han (2012) noted that many computer programs provide approximate average parameter estimates for γ in the case of non-convergence and actually mislead practitioners to interpret the outcome of the calibration as successful calibration. Han (2012) proposes that when items are estimated with a fixed γ parameter that the term *random chance parameter* be used since this will lead to a more appropriate interpretation of the γ parameter.

4. Guessing as a Personal Interaction Parameter rather than an Item Parameter

Unless a test taker is directly asked, it will never really be known whether the test taker guessed the correct answer or knew the correct answer if an item was answered correctly. 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 doesn't know the answer at all, would most likely guess. Note that it is not an item that guesses but rather a person and therefore a guessing parameter needs to be a person interaction parameter and not an item parameter.

5. Personal Probabilities

Consider an item *i* with *k* options (alternatives). The test taker assigns probabilities *p(j)* to each response *r(j)* to indicate the chance of each option being correct, where $0 \leq p(j) \leq 1$ and $\sum_{j=1}^{k} p(j)=1$. The test taker's score on the item is then a function of the responses and the expected score *ES* is a product of the probabilities and the function, i.e.

$$ES = \sum p(j) \times f_j(r_1, r_2, \ldots, r_k) \quad (4)$$

Scoring systems of this class are inexhaustible. Any bounded nonnegative function *f(t)* with *f(1)* = 0 and a bounded derivative for $0 \leq t \leq 1$ can be used to construct a reproducing scoring system. For example, for $k = 2$ the scoring rules for the two options are derived from *f(t)* as

$$S_1(r) = \int_0^r f(t)dt \text{ and } S_2(r) = \int_{1-r}^{1} \frac{t}{1-t} f(t)dt \quad (5)$$

If, for example, *f(t)* = 1 – *t* then the scoring rule $S_{1,2} = r_c - \frac{1}{2}r_c^2$ results from *f(1)* = 0 and a bounded derivative $0 \leq t \leq 1$ where r_c refers to the correct response. If *f(t)* = 1 – t^2, then the function results in a set of scoring rules, one for each option, i.e. $S_1 = -\frac{1}{3}r_c^3$ and $S_2 = \frac{1}{3}r_c^3 - \frac{3}{2}r_c^2 + 2r_c$. The item score thus depends on which one of the options is correct. The maximum score for option 1 is ⅔ and for option 2 it is $\frac{5}{6}$ (Shuford, Albert & Massengill, 1966). Many approaches to assigning probabilities to options in items have been considered (De Finetti, 1970). Instead of using the typical MCQ scoring rule of a score of 1 for a correct answer and a score of 0 for an incorrect answer, a scoring rule has to be devised on the basis of the probability that the test taker assigned to the options. Different scoring rules can be devised to do this in general. For example, a (spherical) scoring rule: Score = $\frac{P_c}{\sqrt{\sum P_o^2}}$ where P_c is the probability assigned to the correct answer and P_o is the probabilities assigned to all the options.

Such a scoring rule has some desired properties such as if a probability of zero is assigned to the correct answer, the score will be a minimum, zero, irrespective of the probabilities assigned to the other options and likewise a probability of one assigned to the correct answer would yield a maximum score of one. Note that the expected score is generally less than the observed score. It can also be verified that this scoring rule will advantage a test taker who states the probabilities more realistic. Also note that a "test-wise" test taker can inflate the item score by eliminating one or more options. Another approach considers probabilities assigned to all options and using a quadratic scoring rule: Score $= 1 + P_c^2 - (1 - P_c)^2 - \sum P_o^2$. The main problem with these scoring rules is that the item score is influenced by how the probabilities are distributed over the distracters. It can, for example, be demonstrated that a test taker who rules out some options, can score lower than a test taker who doesn't, which contradicts the logic that a test taker who can rule out one or more options has some partial knowledge. Thus, although the scoring rules have some desired properties, the item scores depend on how the probabilities are assigned to the different options.

6. Option Probability Theory

In Option Probability Theory (OPT) it is purported that these issues can be overcome by basing the scoring rule only on the probability assigned to the correct option and be independent of the probabilities assigned to the other options (Barnard, 2013a, 2013b,2012). Test takers thus assign personal probabilities *p(j)* to the *k* options of an item *i* so that $0 \leq p(j) \leq 1$ and $\sum p(j) = 1$ and only the *p(j)* assigned to the correct option r_c is considered for scoring so that the score *S* is a function *F* of r_c, i.e. $S=F(r_c)$. Regardless of how the item is scored, the expected score *ES* is maximised as a product of the probabilities and the score (see equation 4), i.e.

$ES = \sum p(j) \times S$ where $S = F(r_c)$ for the item score if r_c is the correct option.

The function *F* defines the scoring rule and since the score *S* is a function of *F* it should be a maximum only if $r = p(j)$ for all *j*. The function *F* should encourage realistic assignment of probabilities to obtain more accurate results. If a test taker is realistic about their knowledge, higher probabilities will generally be assigned to items answered correctly and lower probabilities to items answered incorrectly. This will result in most items with the same distribution of *p(j)* so that the relative frequency "correct" of all options assigned *r* approach *p(j)*.It can be shown mathematically that a logarithmic scoring rule is the only rule which has this property for three or more options.

Function *F* can be derived through partial differentiation under the condition that $\sum p(j) = 1$ using the Lagrange multiplier λ:

$$\frac{\partial[\sum(p(j)F(r)+\lambda(1-\sum r)]}{\partial r} = 0 \quad (6)$$

When $p(j)=r$ where $r = r_c$ and for all *k*. Thus

$$p(k).\frac{\partial F(r)}{\partial r} - \lambda = 0 \quad (7)$$

yielding

$$\frac{\partial F(r)}{\partial r} = \frac{\lambda}{r} \quad (8)$$

For $p(j)=r$ for all *k* and substituting *r* for *p(j)* and λ a constant independent of *r*. It follows that

$$F(r) = A \ln(r) + B \quad (9)$$

for constants A and B. If A and B are chosen so that the score is zero for a uniform distribution, then a test taker will score zero if equal probabilities are assigned to all options, indicating that the answer is not known. Furthermore, these values can be chosen to yield a maximum score of 1 if a test taker assigns a probability of 1 to the correct option (and thus 0 to all other options) of item *i*. In other words $s(i) = 0$ if $p(k) = \frac{1}{k}$ and $s(i) = 1$ if $r_c = p(k) = 1$. Note that probabilities less than $\frac{1}{k}$ assigned to the correct option result in negative scores. In the extreme case where a probability of zero is assigned to the correct option, $s(r) = -\infty$. This can be rectified through a correction (tolerance) parameter *t* where $0 < t < \frac{1}{k}$(Dirkzwager, 1998) which can be varied for items with different numbers of options. In the case where a test taker assigns a maximum probability of 1 to a single option through gambling, i.e. the wrong answer, the expected score is

$$E = \frac{1}{k}s(1) + \frac{k-1}{k}.s(0) \quad (10)$$

With *s*(1)=1 and $s(0) = \frac{\ln t+\ln k}{\ln(1-tk+t)+\ln k} = \frac{f(t)}{g(t)}$.Note that $s(0) = -\infty$ for $t = 0$ and approaches $\frac{-1}{k-1}$ when *t*tends to$\frac{1}{k}$. This means that the adjustment is infinite for zero tolerance and a monotone increasing function of *t* (minus adjustment for non-zero tolerance) approaching $\frac{-1}{k-1}$ for *t* tending to $\frac{1}{k}$.

The maximum tolerance $t = \frac{1}{k}$ for large *k* yields an expected score approaching zero. A test taker can therefore gamble with 100% on the most likely option without adjustment. This is like a test taker gambling on a MCQ – the expected score is zero and is maximised through guessing an option. The smaller the tolerance parameter *t*, the more guessing is discouraged. Test takers sometimes tend to assign high probabilities to incorrect options which would indicate an overestimation of knowledge or otherwise assign low probabilities to correct options which would indicate an underestimation of knowledge. To determine how realistic a test taker has assigned probabilities as a function of their knowledge, a "realism" index can be calculated. When a probability *p*is assigned to a large number *f* of options of different items of which *f(t)*are correct, it is expected that $\frac{f}{f(t)} = r$. A test taker is well calibrated ("realistic") if *r(j) = p(j)* for probabilities *p*. If a linear relationship, e.g. $p = ar + b$, is assumed between *p* and *r*, the probability *p* can be estimated from the reported *r*. For *k* options, summation over options yields$b = \frac{1-a}{k}$for each item. For every value of *r*, the corresponding *p* is estimated as $\frac{f.t(r)}{f(r)}$ and hence, for a least squares estimate of *a*, a minimal function of *a* can be chosen so that $F'(a) = 0$ and*a* can be solved from

$$a = \frac{\sum[ft(r)r-\frac{ft(r)}{k}-\frac{f(r)}{k}+\frac{f(r)}{k^2}]}{\sum[f(r)r^2-2f(r)\frac{r}{k}+\frac{f(r)}{k^2}]} \qquad (11)$$

from which it follows that

$$a = \frac{k\sum\sum t-n}{k\sum\sum r^2-n} \qquad (12)$$

This estimate of *a* yields the best least squares fit when a test taker's probabilities *p* are estimated with the linear formula $p = ar + \frac{1-a}{k}$. If *a* = 1, no correction is necessary since the test taker is as realistic in the probability assignments as possible and no tolerance for overestimation is necessary. If *a*< 1, the test taker is assigning too extreme probabilities, i.e. overestimating knowledge. If the tolerance is set as $t = \frac{1-a}{k}$ the probabilities become more realistic. If *a*> 1, the test taker assigns probabilities too low, i.e. underestimates knowledge in which case the tolerance parameter *t* is negative.

The tolerance parameter can be considered as a measure of "how serious" a test taker's probability assignments are regarded. With $t = \frac{1}{k}$, the test taker's "gambling" is not taken serious and no weight or significance is given to the reported probabilities. If $t > \frac{1}{k}$ or *a*< 0, the test taker assigns too low probabilities to correct options. The hypothesis that a test taker is realistic in assigning probabilities can be tested. A test taker is "perfectly" calibrated when for each value of *p*, the proportion correct of the options this *p* assigned to, is equal to *p*. This hypothesis can be tested as follows: Count, for each *p*, the total number *f(p)* of options *p* assigned to in the test. The expected number correct equals Exp*ft(p)=p.f(p).* Using a chi-squared test, it can be tested if the distribution of the observed *ft(p)* equals the predicted distribution of the expected Exp*ft(p):*

$$\chi^2 = \sum \frac{(ft(p)-Expft(p))^2}{Expft(p)} \qquad (13)$$

Since the chi-square distribution is only sufficiently approximated if Exp*ft(p)*> 5 for all *p*, *ft(p)* and Exp*ft(p)* are summed over increasing intervals of *p* so that $\sum$Exp*ft(p)* $\geq$ 5 in each interval. The degrees of freedom are equal to the number of intervals minus 1. In tests where the emphasis is on a pass/fail decision, the score is the most important result and a realism correction should not be major. But, not correcting for realism would result in possible unrealistic assigning of probabilities and less accurate calibrations. A compromise can be found through modification of the tolerance parameter for each test taker or through modification of the probability assignments.

Modifying the tolerance parameter can penalise realism for underestimation on the one hand and be too lenient for overestimation on the other hand whereas adjusting the probability assignment corrects both under- and overestimation with reward for realism in all cases. The realism score is based on all probability assignments and therefore, if a test taker is overestimating probability assignments, all of them will be reduced and if a test taker is underestimating on average, all probabilities are taken as more extreme. In both overestimation and underestimation the corrected result will be lower than for perfect realism – a desirable outcome. The scoring rule, with a tolerance parameter and correction for realism, yields scores that can be negative. Such scores are (just like in the case of ability estimates derived from Rasch or IRT calibration) not commonly used for reporting purposes. OPT measures therefore have to be converted to a reporting scale that is more "traditional". A stanine scale can, for example be used and performance can be expressed in terms of descriptors for each point on the scale or a dedicated reporting scale can be developed on which certain scores have specific meanings. In deriving the conversion rules, one has to be careful that the conversion is not too lenient for certain values. For example, responding with $\frac{1}{k}$ to all options of all items in a test will yield a "raw" score of zero, which can convert to a score just below a cut-score of "50%". A better conversion would be a monotonic increasing function which is concave up to the cut-score and convex above the cut-score. Although such a function will lose some properties of a "traditional" S-shaped curve, it will not affect the cut-score and will have the advantage that it will yield a larger increase in the reported score for smaller increments in the raw OPT measures.

7. Rationale for OPT Scoring

In a four-choice item a test taker has a 25% chance of randomly choosing the correct answer. If the correct answer is chosen, the test taker scores one and otherwise scores zero in "traditional scoring". If the test taker can rule out two of the four options, there is a 50/50 chance to randomly select the correct answer. Again, if the correct answer is selected the score is one or otherwise zero. But, if the test taker is more certain about one of the two remaining options, a 60/40 assignment cannot be implemented in a dichotomous scoring system. In OPT this is possible and more credit will be given if 60% was assigned to the correct option, than if 40% was assigned to it – allowing for partial knowledge. Extending this argument the scoring function can be defined to assign a maximum number of points to a correct response with 100% certainty. The logarithmic scoring rule allocates fewer points as the level of certainty decreases and reaches zero at 25%. In other words no points are allocated for uncertainty where the same chance is allocated to the four options. However, if less than chance is assigned to the correct answer, less than zero points should be awarded to the response. For such assignments the scoring function is more severe and the negative points increase exponentially as the percentage assignment to the correct option decreases. The negative score per item can be used to flag the "severity" of a misconception, etc.

8. Examples

Consider a hypothetical 10-item test with four-choice questions and suppose that the scoring function is defined in such a way that random guessing results in a score of zero and 100% assigned to the correct answer results in a score of 1. Further suppose that the reporting function converts scores to a scale from 0 to 10 on which 5.5 is the cut-score. If 70% is assigned to the correct answer of five items and 40% is assigned to the correct answers of the remaining five items, the resulting score is 7.53 with a realism index of 2.98 which can be interpreted as a well above average performance and realistic assignment of probabilities.(Perfect realism is at 0; positive values indicate underestimation of knowledge and negative values indicate overestimation of knowledge.) In this case 10% was allocated to each of the remaining incorrect options for the first five items and 20% for the last five items. If two incorrect options are ruled out in all ten items and 30% is allocated to the only incorrect option for the first five items and 60% for the last five items, the score remains 7.53 but the realism index changes to -0.18, giving credit for the elimination of two options in each item.

A 50% assignment to the correct response of all ten items results in a score of 7.33 and realism index of 2.97. This can be interpreted as the test taker ruling out two options in each item and allocating too low probabilities to the correct answers. A 20% assignment to the correct answers of all ten items (5% less than random chance) results in a score of 4.51 which is below the cut-score of 5.5 with a realism index of -3.87 which indicates that the test taker assigned too low probabilities to the correct answers. The score drastically decreases to 3.1 if 10% is assigned to the correct answers of all ten items together with a realism index of -9.46 indicating that significantly low percentages were assigned to the correct answers.

9. Real Cases: An Exploratrory Study

A small study was designed to investigate the implementation of OPT. Some 89 Year 5 students participated in the study in which 20 five-option mathematics MCQs were used in two parallel tests. The cohort was randomly divided into two groups and the one group was first given a conventional version of the test whilst the other group was first administered the OPT version of the test. The OPT version was administered through EPEC's software (EPEC prognoser©, 2013). After a break the first group was administered the OPT version whilst the second group sat the conventional test. In addition to counterbalancing, the two tests were constructed from the same blueprint and item information functions were used to ensure that they were as equivalent as possible. Before administering the OPT version students were briefly instructed to assign percentages to the options in each item according to their beliefs of the correct options. All students completed each test in 30 minutes. The conventional test was scored "traditionally" by allocating one mark for each correct answer and zero to an incorrect answer. The OPT version was scored according to the scoring function and converted to a 0 to 20 scale for more direct comparisons. A realism index was computed for each student in OPT.

9.1 Results

The students generally performed better in the conventional version than in the OPT version. An independent samples t-test was conducted to compare the scores on the two versions of the test. There was a statistically significant difference in scores on the conventional version ($M = 11.75$, $SD = 3.65$) and the OPT version ($M = 8.40$, $SD = 3.50$; $t(176) = 6.25$, $p = 0.00$). The magnitude of the difference in the means was large (eta squared = 0.18). Scores in the conventional version ranged from 2 to 18 whilst scores in the OPT version ranged from 2.86 to 17.06. There was a significant difference in performance in favour of the 54 boys to the 35 girls in both the conventional and the OPT versions of the test. The magnitude of the differences in means was large in both cases – eta squared of 0.22 in the conventional test compared to 0.17 in the OPT version. Practice effects were looked at by comparing the performance of the 45 students in group 1 who sat the conventional test first to the 44 students in group 2 who sat the OPT version first. For the conventional test, group 2 ($M = 12.32$, $SD = 3.28$) outperformed group 1 ($M = 11.20$, $SD = 3.94$; $t(87) = 1.45$, $p = 0.15$). The difference in means was not significant and the magnitude of the difference was small (eta squared = 0.02). In the OPT version, group 1 ($M = 8.44$, $SD = 3.64$) outperformed group 2 ($M = 8.26$, $SD = 3.44$; $t(87) = 0.23$, $p = 0.82$). The difference in means was not significant and the magnitude of the difference was very small (eta squared = 0.00). These results suggest that there were no significant practice effects.

The OPT version was also scored in a different way. Based on the assumption that students were likely to answer the item correctly if only one option had to be chosen (like in conventional scoring) if they assigned 60% or higher to the correct option; likely to answer the item incorrectly if they assigned 40% or less to the correct option; and scored half (rounded up) for items assigned between 40% and 60% to the correct option, whole number scores were calculated for each student. For example, if a student assigned 60% or higher to the correct option of eight items, 40% or less to the correct option of nine items and between 40% band 60% to the correct option of the remaining three items, the student's score is 8 + 0 + 2 = 10. A one-way between-groups analysis of variance was conducted where group 1 scores were the OPT scores, group 2 scores were the dichotomised OPT assignments and group 3 were the scores on the conventional test. There was a statistically significant difference at the $p < 0.001$ level [$F(2, 264) = 20.98$, $p = 0.001$]. The effect size, calculated using eta squared, was 0.14 which can be considered as a large effect. Post-hoc comparisons using the Tukey HSD test indicated that the mean score for group 1 ($M = 8.40$, $SD = 3.50$) was significantly different from group 2 ($M = 10.70$, $SD = 3.42$) and from group 3 ($M = 11.75$, $SD = 3.65$) but that the difference between groups 2 and 3 was not significant. These results are not surprising as the correlation between group 1 and group 3 [$r = 0.55$, $n = 89$, $p < 0.001$] was much lower than the correlation between group 1 and group 2 [$r = 0.79$, $n = 89$, $p < 0.001$] and between group 2 and group 3 [$r = 0.71$, $n = 89$, $p < 0.001$].

9.2 Discussion of Results

If OPT assignments are dichotomised, the scores were not significantly different from the conventional scores. This result confirms the equivalence of the two tests used and also that a probability format can be used whenever a conventional format is used.

However, the OPT format permits additional information about guessing, understanding and misconceptions. One student, "Fiona", answered 17 of the 20 items in the conventional test correctly and thus scored 85%. In the OPT version, she was quite realistic about her knowledge and generally assigned realistic probabilities to the options in the items. She assigned a high percentage (95%) to an incorrect option of only one item. (She assigned 0% to the correct answer.) For this item she overestimated her knowledge most likely due to a misconception. She was not confident about the answers of six of the 20 items and assigned relatively low percentages to the correct answers of these items. Her responses suggested that she did allow some possibility that the correct answers were correct – unlike the item in which she was 95% sure that an incorrect answer was correct. The OPT scoring function allowed some score for this, depending on the percentages assigned. She was unsure about two items and left all the options at their default equal percentages. When confident about answers, she assigned high percentages. She assigned high percentages to 11 items answered correctly - 100% to the correct answers of three of these items, 95% to two of them and 90% to six of these items. Her OPT score converted to a score of 69.6% which can be interpreted as around 14 out of 20. (Her dichotomised OPT score was 12.) When comparing the OPT score to the conventional score, it should be noted that she answered the parallel item to which she assigned 95% to an incorrect option incorrectly and omitted the two items linked to those which she left at default percentages in OPT. In the conventional version, she responded correctly to the parallel versions of three of the six items that she was unsure about in the OPT version and incorrectly to the other three items. Assigning lower percentages to the correct options of six items in OPT resulted in lower credit than was given in the conventional version. The lower percentages assigned can be interpreted as less confident about the answers and it can be suspected that she guessed, perhaps some educated, in the conventional version of these six items – answering three correct.

Her realism index of -1.95 indicated that she was quite realistic about her knowledge. (An ideal value is zero.) She was penalised significantly for assigning 95% to the incorrect option of one item, but the realism index minimised with realistic assignment of probabilities to the other 19 items. She underestimated her knowledge, especially in the 11 items answered correctly and overestimated her knowledge mainly in the one item answered incorrectly. The assignment of 95% to the incorrect option of one item, came at a severe cost. If equal default probabilities were assigned to each option, her OPT score would have been 78% and the realism index 0.07. Of course, if she answered the item correctly, her OPT score would be 81% and the realism index 0.09. If she answered the parallel item in the conventional test incorrectly, her score would be 80% which compares very well with her OPT score. This means that the misconception had a significant impact on her results. Another student, "Sam", answered seven of the 20 items in the conventional version correctly and thus scored 35%. His OPT score of 36.1% was marginally higher, but the realism index of -20.92 indicates some unrealistic, aberrant responses.

Sam incorrectly assigned 100% to incorrect answers of six items resulting in negative OPT scores and contributing significantly to a high negative realism index. He assigned 20% to the correct answer of two items and since the items were all five-choice, scored zero on these two items. He allowed equal probabilities (default values) to three items, also resulting in item scores of zero, and was confident about the answers of nine items answered correctly. In the OPT scoring, a score of zero was assigned to five items, a negative score to seven items and positive scores to nine items resulting in a converted score of 36.1%, i.e. "equivalent to around seven out of 20. When comparing the items pair-wise it was found that five of the items answered correctly in the conventional version were also allocated high probabilities in the OPT version. Sam assigned 50% to the correct option of an item in the OPT version which he responded correctly to in the parallel item in the conventional version and only 20% to another such item. It thus appears that he most likely guessed these two items correctly in the conventional version and risked high probabilities to more items in the OPT version.

10. Conclusions

Guessing answers in MCQs has always been an issue. Whereas the correction for guessing formula is difficult (if possible) to defend, modern test theories deal with guessing through fit statistics and/or a guessing parameter. Since people, rather than items, guess it is argued in OPT that guessing should be dealt with as a person-item interaction parameter. Results from the relatively small study suggest that there is a high correlation between scores obtained conventionally and through OPT. However, one significant advantage of the latter is that specific information about guessing is obtained through analysis of certainty about answers which increases reliability (Rippey, 1970). OPT can be used to illuminate misconceptions and their nature, where there is uncertainty and where a test taker is certain and correct or incorrect.

In traditional scoring there is no information of whether the test taker knew or guessed a correct answer. Having more information about responses can increase robustness of decisions about borderline test takers. For example, a borderline candidate who has assigned a high percentage to an incorrect answer is less likely to check afterwards whether the response was correct than a candidate who assigned say 60% to the correct answer. Yule (1988), for example, found that hearing-impaired students who were very confident in their identification of wrong responses were less likely to improve with teaching. A major limitation of the findings in the study was that students were not instructed sufficiently in the logic of the scoring in OPT. This resulted in students taking too large risks with direct implications in the scoring and especially the realism index. Only 12 of the 89 students scored higher in OPT than in the conventional test. More modest assignment of probabilities to answers that students were less sure of would have resulted in more realistic responses. Comparing items pair-wise between the conventional mode and the OPT mode suggest that test takers do indeed make educated guesses when items are scored dichotomously. In OPT, test takers who take risks by assigning high probabilities to incorrect options lower their overall scores and usually score less than their conventional scores whilst test takers who assign moderately and more realistic probabilities often have slightly higher scores in the OPT version than in the conventional mode.

11. References

Andrich, D. (1988). Rasch models for measurement. Newbury Park, CA: Sage.

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

Barnard, J.J. (2013a). Die bepunting van meerkeusigevrae. Suid-Afrikaanse Tydskrifvir Natuurwetenskappeen Tegnologie 32(1), 7pp.

Barnard, J.J. (2013b). Option Probability Theory. Available from http://www.epecat.com/EPEC_Option_Probability_Theory.

Bond, T.G. & Fox, C.M. (2007): Applying the Rasch model. Second edition. Lawrence Erlbaum. London.

Crocker, L.M., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston Inc.

De Finetti, B. (1970). Logical foundations and measurement of subjective probability. ActaPsychologica, 34, 129-145.

Dirkzwager, A. (1998). Personal communication.

Ebel, R.L., &Frisbie, D.A. (1991). Essentials of educational measurement. New Jersey: Prentice Hall.

EPECPrognoser [Computer software] (2013). www.EPECprognoser.com Melbourne: Excel Psychological and Educational Consultancy (EPEC).

Frary, R.B. (1988). Formula scoring of multiple-choice tests (correction for guessing). Educational Measurement: Issues and Practice. 17(2): 33-38.

Haladyna, T.M. (2004). Developing and validating multiple-choice test items. London: Lawrence Erlbaum.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage Publications.

Hambleton, R.K., &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.

Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2): 149-174.

Phelps, R.P. (2000). Trends in large-scale testing outside the United States. Educational Measurement: Issues and Practice, 19(1): 11-21.

Rippey, R. (1970). A comparison of five different scoring functions for confidence tests. Journal of Educational Measurement, 7, 165-170.

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

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

Shuford, E.H. (Jr)., Albert, A., &Massengill, H.E. (1966). Admissible probability measurement procedures. Psychometrika, 31, 125-145.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-166.

Yule, G. (1988). Highly confidence wrong answering and how to detect it. ELT Journal, 42(2), 84-88.