ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 11, Issue, 08, pp.6202-6206, August, 2019

DOI: https://doi.org/10.24941/ijcr.36208.08.2019

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

MEASUREMENT ERROR IN ASSESSMENT

***John J. Barnard**

EPEC Pty Ltd. / University of Sydney, Faculty of Medicine and Health, Australia

ARTICLE INFO

Article History:
Received 24th May, 2019
Received in revised form
20th June, 2019
Accepted 28th July, 2019
Published online 31st August, 2019

Key Words:

Measurement error, Rasch,
IRT, OPT, guessing.

****Corresponding author:*** **John J. Barnard**

ABSTRACT

Multiple-choice tests are usually scored dichotomously, i.e. as correct or as incorrect. A correct response is scored one and an incorrect response is scored zero. Using these scores different psychometric paradigms and models are used to analyse the data and to quantify the performance of the test takers. Classical Test Theory commonly add the question scores to obtain a total score whilst Rasch and Item Response Theory models estimate measures from probabilities. In this paper an argument is made that dichotomous scoring includes significant measurement error as uncertainty of responses is not considered. It is demonstrated how Option Probability Theory can overcome this through assigning percentages to one or more options according to the test taker's mental processes.

***Citation: John J. Barnard*, 2019.** "Measurement error in assessment", *International Journal of Current Research*, 11, (08), 6202-6206.

INTRODUCTION

A measurement theory can be described as a set of assumptions and definitions from which psychometric properties of measures can be determined (Allen and Yen, 1979). Most definitions of measurement relate to the assignment of numbers to objects (Lord and Novick, 1968). Whereas some definitions of measurement are limited to the assignment of numbers in defined ways, others go further and require that the numbers represent the distances between objects on a continuous line in terms of what is being measured (Wright and Stone, 1979). Classical Test Theory (CTT), also referred to as true score measurement theory, assume that a test taker's observed score (raw score) on a test is comprised of a true score and an uncorrelated measurement error (Crocker and Algina, 1986). A true score is hypothesised as the arithmetic mean of the distribution of scores obtained by a test taker in independent repeated measures and the error of measurement (error score) is the difference between an observed score and its theoretical true score counterpart (Harvill, 1991). The error score is that part of the observed score which is unsystematic, random and due to chance. Reliability is a key concept in CTT and is operationalised through the Kuder-Richardson Formula 20 (KR-20) for dichotomously scored multiple-choice questions. The KR-20 formula is the result of two factors, namely item variance and test variance. As such it is directly proportional to the variances of the test, i.e. the sum of the item variance remains constant and as the test variance increases, so does the reliability. As such reliability paints an incomplete picture. The standard error of measurement (SEM) is defined as the standard deviation of errors of measurement that is associated with the test scores for a specific group of test takers, i.e. it is a measure of the variability of the errors of measurement and is directly related to the error score variance. The SEM is directly related to reliability, i.e. the ratio of the true score variance to the observed score variance, and can be calculated[1] as

$$SEM = S_X\sqrt{(1 - r_{xx'})} \tag{1}$$

where S_X is the standard deviation of observed scores and $r_{xx'}$ is the reliability of the test (Peterson, 1994). Only if the reliability of the test is a perfect 1, will the SEM be zero, i.e. no error of measurement. The unit of measurement for the SEM is the same as the unit of measurement of the original test scores and allows for statements about the precision of test scores of test takers. Interpretation of the SEM is based on a normal distribution. By adding multiples of the SEM to an observed score, the precision or confidence of the score can be expressed – the higher the precision, the wider the score band. In the Standards for Educational and Psychological Testing (1985) it is defined that the standard deviation of errors of measurement is associated with the test scores for a specified group of test takers, i.e. the SEM is a test characteristic estimated from a specific group of test takers. Although the SEM is usually reported as a single value, Harvill (1991) recommends that SEMs at different score levels be used in calculating score bands rather than a single SEM value and also noted that the type of reliability coefficient used in

[1] A good approximation for the SEM is 0.432 times the square root of the numbers of items in a test.

calculating the SEM can make a difference, both computationally and logically, based on the research of Feldt, Steffan and Gupta (1985). In Rasch/Item Response Theory (IRT) models, the estimates of item parameters are independent of the distribution of the trait in the sample being tested and test takers' scores do not depend on the particular sample of items administered. These models assume local independence, i.e. that responses to questions (items) are independent for test takers at the same level of the latent trait being measured. The differences in assumptions between CTT and Rasch/IRT result in different procedures to estimate psychometric properties – reliability in particular. Rasch/IRT models go beyond the set of items in a test. Whereas CTT requires a new set of descriptions for a different test since there is no direct relationship between item scores and test takers' scores, Rasch/IRT uses mathematical models for predicting the probability of success of a test taker on an item, depending on the test taker's ability and the item difficulty (Wu and Adams, 2007). A transformation to raw scores is applied so that distances between the locations of two test takers is preserved, independent of the particular items administered. This allows for exchanging of items in a test with other items that measure the same latent trait – something that cannot be done in CTT since CTT doesn't make any assumptions about the latent trait.

The Rasch/IRT probability function, referred to as an item characteristic curve or item response function, has the general mathematical form

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

where person *n* with ability β has a certain probability *P* to respond correctly to item *i* with difficulty δ, given *e* the base of the natural logarithm, α is the discrimination parameter, γ is the guessing (pseudo chance) parameter and *D* is a scaling constant in IRT. For the 2-parameter IRT model, γ is zero and for the Rasch/1-parameter IRT model, α is one in addition to γ taken as zero.

A first step in Rasch analysis is to determine if the data fit the model (Wright and Stone, 1979). This is a measure of accuracy which is quantified by fit statistics. Residual based weighted (Infit) and unweighted (Outfit) mean squares are commonly used in the form of χ^2 variates where *X* is the observed score and *E* is the expected score:

Outfit Mean Square Infit Mean Square

$$\chi^2=\sum\frac{(X-E)^2}{Variance}\ \chi^2=\frac{\sum(X-E)^2}{\sum Variance} \qquad (3)$$

Fit mean square values have an expectation of one and values less than one are interpreted as "over" fit and values greater than one are interpreted as "under" fit. The outfit statistic is more sensitive to outliers than the robust infit statistic which is more related to discrimination. Outfit is commonly used for diagnostic purposes, to find aberrant response patterns.

Precision on the other hand is quantified by standard errors and can be related to the concept of reliability (Wu and Adams, 2007). Every Rasch/IRT measure has an associated standard error and is calculated as one divided by the square root of the statistical information in a measure, i.e.

$$SEM=\frac{1}{\sqrt{Information}}=\frac{1}{\sqrt{I(\theta)}} \qquad (4)$$

Conceptually information is given by the slope squared divided by the conditional variance, i.e.

$$I(\theta)=\frac{(P(\theta)')^2}{P(\theta)(1-P(\theta))} \qquad (5)$$

where $P(\theta)$ is the probability of a correct answer. The computational equation for the 3-Parameter Logistic IRT model (3PL) is given by

$$I(\theta)=D^2a^2\frac{Q(\theta)}{P(\theta)}\left(\frac{P(\theta)-C}{1-C}\right)^2 \qquad (6)$$

Note that (6) becomes

$$I(\theta)=P(\theta)Q(\theta) \qquad (7)$$

for the 1-Parameter Logistic IRT (1PL) and Rasch models where D=1; a=1 and C=0, i.e. item information is equal to item variance.

Information functions are used to calculate the Standard Error of the Estimate (SEE), an IRT statistic which is interpreted in the same way as the SEM of CTT. In Rasch measurement an estimate's standard error is the modelled standard deviation of the normal distribution of the observed estimate around its true value (Linacre, 1994). For example, a standard error of 0.385 yields 99% confidence that the true estimate is within one logit from the reported estimate. For N well targeted observations, the minimum SE is $\sqrt{\frac{4}{N}}$ and the maximum SE is $\sqrt{\frac{9}{N}}$ so that for 51 observations the minimum SE is 0.28 and the maximum is 0.42. Precision of ability measures can be increased by increasing the test length and precision of item difficulty measures can be increased by increasing the sample size.

Measuring test takers' abilities

The main goal of measurement is to quantify test takers' performance. This is achieved through the building blocks of a test, namely the questions (items). The items are used to derive a score which is used to express performance as a number on a scale. In CTT it is common practice to use the raw (number correct) total score as such an indicator. It is acknowledged that the total score includes some error resulting in reporting the (fixed) SEM. Rasch/IRT overcomes the restrictions of sample dependency through stochastic processes and calculating probabilities of a correct response to obtain ability measures on an interval level scale. It doesn't matter whether CTT, Rasch or IRT is used as the underlying measurement paradigm, the scores/measures always have associated errors. Traub and Rowley (1991) mention that there are many sources for errors in tests scores, and in multiple-choice questions (MCQs) guessing is arguably one of the most prominent. Whereas a correction for guessing was introduced in CTT (Hogan, 2003), Rasch modelling uses fit statistics to identify aberrant responses and response patterns which may indicate guessing. A correct response to an item with difficulty greater than a test taker's ability is considered as an unexpected response and should be flagged if the item is answered correctly. The 3PL IRT model incorporates a pseudo chance parameter which is often interpreted as indicative of guessing. Whether CTT, Rasch or IRT is used, MCQs are usually scored dichotomously, i.e. as either correct or incorrect and scored as either 1 or 0. The stochastic approach of Rasch/IRT computing the probability of a test taker to answer an item correctly or incorrectly remains based on a 0/1 score. A test taker may have decided that the correct answer is, say, one of two options and since only one option can be chosen, the test taker has to choose one of these two options. If the test taker had to make a choice between the two options and weigh them as a 50/50 chance of being correct, there is significant error in both whether the correct or the incorrect option was finally chosen as the answer, if one of the two options was correct. The same rationale holds for other response combinations. A dichotomous 0/1 item score therefore doesn't account for a possible large amount of error. When considering the answer to an MCQ, a test taker would normally weigh all the options and choose the option most likely correct in their opinion, even if it is an educated guess. If an incorrect option is chosen as the answer, the test taker scores 0. This score does not reflect the extent to which the correct option was considered. To quantify this, Bayesian's rule can be applied: Each item entails k hypotheses for each option i the hypothesis H_i that i is correct. The initial prior hypothesis is that H_i are all equal, i.e. $p(H_i) = \frac{1}{k}$ where k is the number of options. Thus, the conditional probabilities are $H_i \backslash M$ where M is the mental process. The test taker chooses the option that is correct in their opinion, i.e. $p(M\backslash H_i)$. According to Bayes' rule $p(H_i\backslash M) = \frac{p(H_i).p(M\backslash H_i)}{p(M)}$. The test taker thus selects the option for which $p(H_i\backslash M)$ is a maximum. But this option is not necessarily the correct option and therefore $p(H_i\backslash C)$ is a better measure of M since it is proportional to $p(M\backslash H_i)$, the ability to choose the correct option C. This means that the item score is equal to the probability assigned to the correct option and not the option to which the maximum probability is assigned. This is in sharp contrast with 0/1 scoring where the test taker chooses the option that is correct in their view, even if it is a guess. This score should be a monotonic increasing function of $In\,p(H_i\backslash M)$ for the correct option i which can be achieved with a linear function of$\ln p(H_i\backslash M)$, i.e. the scoring rule should be a logarithmic rule.

To derive the scoring rule, consider item i with k options. The test taker assigns probabilities to each option so that the response vector $\bar{r} = (r_1, r_2, \dots, r_k)$ where $0 \leq r_i \leq 1$ and $\sum_{i=1}^{k} r_i = 1$. The score s_i is a monotonic increasing function of response vector $\bar{r}$, i.e. $s(r_i) = F(r_i)$. Irrespective of the scoring rule, the test taker must estimate the likelihood to maximise the expected score E. Such a scoring system where $\bar{r} = \bar{p}$ maximises if $E = \sum p(i) \times F(r_i) = \sum p(i) \times s(r_i)$. Scoring systems of this class are inexhaustible. If it is based on the probability assigned to the correct option only, then $s_i = F(r_c)$ for $0 \leq p_i \leq 1$ and $\sum_{i=1}^{k} p_i = 1$. The expected score E can be written as a linear function $E = \sum p(H_i\backslash M).\log p(H_i\backslash M)$. By substituting $p(H_i\backslash M)$ with pi and considering only the response to the correct option, s can be maximised: $E = \sum p_i.s = \sum p_i F(r_i)$.Thus,$s$ is a function for F of r_i and the expected value is a maximum if and only if $r_i = p_i$ for all i. Function F can be derived through partial differentiation under the condition that $\sum r_i = 1$ using the Lagrange multiplier λ:

$$\frac{\partial[\sum(p(j)F(r)+\lambda(1-\sum \mathrm{r})]}{\partial r} = 0 \qquad (8)$$

when $p(j)=r$ where $r = r_c$ and for all k. Thus $p(k).\frac{\partial F(r)}{\partial r} - \lambda = 0$ yielding

$\frac{\partial F(r)}{\partial r} = \frac{\lambda}{r}$ for $p(j)=r$ for all k and substituting r for $p(j)$ and λ a constant independent of r. It follows that $F(r) = A\ln(r) + B$ for constants A and B. For scoring purposes, A and B can be chosen so that $s_i = 0$ if $p_k = \frac{1}{k}$ which indicates that the test taker doesn't know (guesses) and $s_i = 1$ if $r_c = p_k = 1$.

This scoring rule estimates the probability assigned to the correct option as a measure of "true" ability and not the probability to answer the question correctly as in Rasch/IRT which is a measure of confidence that the option is correct. The difference is subtle but significant. Barnard (2015) labelled this approach Option Probability Theory (OPT) and demonstrated how this approach minimises uncertainty and illuminates guessing. Unlike dichotomous scoring, a test taker assigns a percentage to any number of answer options and the percentage assigned to the correct option is scored by means of the logarithmic scoring function $\Psi = F(r) = A\ln(r) + B$ which penalises low percentages assigned to correct answers (higher percentages assigned to incorrect answers). The theory was applied in a study comparing the performance across three assessment formats in online test administration; Multiple Choice Questions (MCQs), Short Answer Questions (SAQs) and Option Probability Theory (OPTs) using parallel tests on a sample of 276 Bachelor of Medicine and Bachelor of Surgery (MBBS) students. (Trigg, Barnard, Devitt and Pham, 2016). The authors found that the students performed best in the MCQ and worst in the OPT and concluded that the OPT scores purport to provide information about individuals' guessing which cannot be captured through MCQs and the lower OPT scores were due to penalising for uncertainty of correct answers.

Comparing ability estimates

To quantify the difference between dichotomous scoring and OPT scoring, 51 students were administered a 40-item four-choice practice medical exam online. All students responded to all items and the data was analysed within the Classical Test theory (CTT), Rasch Measurement Theory and Option Probability Theory frameworks. For the dichotomous scoring, a(classical) mean of 23.73 (59.33%) with a standard deviation of 3.77 (9.43%) was obtained. The relatively low alpha (KR-20) reliability of 0.48 had an associated SEM of 2.72 (6.80%). This means, for example, that there is a 68% certainty that a student who scored 30 (75%) has a "true ability" within the range [27.28; 32.72], i.e. [68.2%; 81.8%]. The Rasch calibration yielded a mean item difficulty of 0 logits (SD 1.17), due to standardising on item difficulty, and an item estimate reliability of 0.90. The students had a mean ability of 0.59 logits (SD 0.49 logits) and a mean estimate reliability of 0.41. Using a score equivalence table, it is found, for example, that a student with a score of 30 (75.0%) has an ability estimate of 1.40 and a standard error of 0.40 logits. A SE of 0.40 is associated with a (classical) reliability index of approximately 0.84 and if this value is substituted in equation 1 using the SD of 9.43% the resulting SE is 3.77% which is less than the CTT SE of 6.80%. The Rasch score is thus more precise than the

CTT score. Table 1 shows the SE range for these scores, showing the more precise Rasch range.

Table 1. SE confidence ranges for CTT and Rasch scores

	Range %
CTT	[68.2; 81.8]
Rasch	[71.2; 78.8]

In the OPT scoring, an estimated score, ε, was calculated from the percentages assigned by the students. This estimated score is intended to mimic dichotomous scoring and interprets higher percentages assigned to correct answers as correct responses and low percentages assigned to correct answers as incorrect responses. If 50% was assigned to a correct option, the item is scored as correct (1) and if this happens a second time the item is scored as incorrect (0), repeating this scoring regime for each student's response vector. The estimated score ε thus dichotomises the responses to yield response vectors that mimic MCQ dichotomous scoring. It is expected that ε will be greater than Ψ for test takers who don't have a perfect score and that the difference between ε and Ψ will be greater as a function of the percentages assigned and also of the number of questions to which higher percentages are assigned to incorrect answers. This is shown in Table 2.

Table 2. OPT data for 8 possible cases for ε as 75%

	A	B	C	D	E	F	Ψ	ε	Diff	E+F
Case 1	0	0	10	0	10	20	67.9	75.0	7.1	75.0
Case 2	5	5	0	0	10	20	66.4	75.0	8.6	75.0
Case 3	10	0	0	0	10	20	66.0	75.0	9.0	75.0
Case 4	0	0	0	10	5	20	65.3	75.5	10.2	*62.5*
Case 5	0	0	10	0	15	15	63.4	75.0	11.6	75.0
Case 6	0	0	0	20	0	20	62.8	75.0	12.2	*50.0*
Case 7	5	5	0	0	15	15	61.9	75.0	13.1	75.0
Case 8	10	0	0	0	15	15	61.5	75.0	13.5	75.0

where
A = Number of items to which a high probability is assigned to an incorrect option.
B = Number of items to which a moderate probability is assigned to an incorrect option.
C = Number of items to which no preference was given to any option, i.e. equal probabilities to all options.
D = Number of items to which a low probability is assigned to a correct option.
E = Number of items to which a moderate probability is assigned to a correct option.
F = Number of items to which a high probability was assigned to a correct option.
Ψ = OPT scaled score
ε = Estimated OPT score
Diff = ε – Ψ

Note that ε = E+F for all cases except for cases 4 and 6 where credit was given to the ε score from items in column D where 50% was assigned to the correct answer of the items. To a lesser extent than the items in column E, this can be interpreted as some partial knowledge. This difference is a maximum for Case 6 where there were the most items in column D, followed by Case 4.

The difference between ε and Ψ is the smallest for Case 1. In this case 100% was assigned to the correct answer of 20 items; 80% to the correct answers of 10 items and equal percentages (20%) to all of the options of 10 items. The Ψ score is 7.1% less than the ε score because of the 20% uncertainty of the 10 items in column E. (The 10 items in column C did not attract any penalty because no preference was given to any option through assigning 20% to each option which indicates that there was no guessing in these items). The difference between the ε score and the Ψ score increases from Case 1 to Case 8 almost exclusively as a function of the number of items in column F where high percentages were assigned to the correct answers. The exception was Case 5 where E + F was greater in Case 5 than in Case 6 due to the 20 items in column D in Case 6. In all cases ε = 75% and greater than the Ψ score due to the penalties for the uncertainties. The Ψ scores ranged from 61.5% to 67.9%, i.e. penalties ranging from 7.1% to 13.5% and thus lower than the lower bounds in Table 1. Thus, although the Rasch estimate was more precise than the CTT score, it did not include the Ψ score. Although the (Rasch) ability estimates for cases 1 and 8 were the same due to the sufficient statistic property of the Rasch model, the Rasch fit index for Case 1 indicated a much less aberrant response pattern than the fit index for Case 8. In sharp contrast, the OPT Ψ score reflects the higher uncertainty in Student 8's score directly.

Summary and conclusion

Different psychometric paradigms can be used to analyse data and to score test takers. Dichotomous scoring of multiple-choice questions is the most popular scoring regime. However, scoring a question as 0 or 1 does not account for large amounts of uncertainty and guessing. Rasch and IRT models attempt to quantify this in different ways, but are based on probabilities assigned to the ability to choose the correct response amongst a number of responses. In contrast the scoring rule in OPT estimates the probability assigned to the correct option as a measure of "true" ability and not the probability to answer the question correctly as in Rasch/IRT which is a measure of confidence that the option is correct. The difference in scores obtained from CTT, Rasch and OPT was demonstrated and it can be concluded that OPT scoring is superior to dichotomous scoring in the sense that it captures uncertainty and guessing and thereby minimising measurement error. Conflict of interest and funding statement: The author declares that there is no conflict of interest and no funding obtained to write and submit this original paper.

REFERENCES

Barnard, JJ. 2015. Option Probability Theory: A quest for better measures. *American review of mathematics and statistics,* 3 (1): 61-69.

AERA, APA and NCME. 1985. Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Allen, MJ., Yen, WM. 1979. Introduction to measurement theory. Monterey, California: Brooks/Cole Publishing Company.

Crocker, LM. and Algina, J. 1986. Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston Inc.

Feldt, LS., Steffan, M. and Gupta, NC. 1985. A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement,* 9: 351-361.

Harvill, LM. 1991. Standard error of measurement. Educational Measurement: Issues and practice 10 (2).

Hogan, TP. 2003. Psychological testing: A practical introduction. John Wiley and Sons Inc.

Linacre, JM. 1994. Sample size and item calibration stability. Rasch measurement transactions, 7 (4).

Lord, FM., Novick, MR. 1968. Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Peterson, RA. 1994. A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research***,** 21 (2): 381–391.

Traub, RE. Rowley, GL. 1991. Understanding reliability. Educational Measurement: Issues and practice 10 (1).

Trigg, M., Barnard, JJ., Devitt, P. and Pham, H. 2016. Comparing examination performance across three online assessment formats in Australian medical students. *International Journal of Research and Current Development.* Vol 1 (1): 11-16.

Wright, BD., Stone, MH. 1979. Best test design. Chicago, IL: Mesa Press.

Wu, M., Adams, R. 2007. Applying the Rasch model to psycho-social measurement: A practical approach. Educational Measurement Solutions, Melbourne.

****