Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 23 Number 14, October 2018 ISSN 1531-7714

From Simulation to Implementation: Two CAT Case Studies

John J. Barnard, *EPEC Pty Ltd / University of Sydney*

Measurement specialists strive to shorten assessment time without compromising precision of scores. Computerized Adaptive Testing (CAT) has rapidly gained ground over the past decades to fulfill this goal. However, parameters for implementation of CATs need to be explored in simulations before implementation so that it can be determined whether expectations can be met. CATs can become costly if trial-and-error strategies are followed and especially if constraints are included in the algorithms, simulations can save time and money. In this study it was found that for both a multiple-choice question test and a rating scale questionnaire, simulations not only predicted outcomes for CATs very well, but also illustrated the efficiency of CATs when compared to fixed length tests.

Obtaining precise scores efficiently is one of the main goals of assessment. Computerised Adaptive Testing (CAT) purports to be an optimal mode of assessment to achieve this goal. A computerized adaptive test (CAT) is a test administered by computer that dynamically adjusts itself to the trait level of each test taker as the test is being administered. By tailoring the testing through intelligent question selection, CATs can reduce test length by at least 50% without compromising measurement precision (Barnard, 2015; Wagner-Menghin & Masters, 2013; Weiss, 2011).

Most CAT programs focus on achievement testing using dichotomously-scored multiple-choice questions (MCQs). Each question has one difficulty value (threshold) which is used to determine which question needs to be administered next. Since the ability of the respondent and the difficulty of each question are located on a common scale, the most appropriate question to administer next can be determined from the respondent's current ability, estimated from previous responses (Van der Linden & Glas, 2003; Wang & Vispoel, 1998; Weiss, 1982). In contrast to dichotomous MCQs, questionnaires are usually polytomous with multiple thresholds as each question has a number of possible response categories.

CATs are commonly based on Rasch measurement or Item Response Theory (IRT) which locates the measure of each score on a common interval scale. Whilst Rasch/IRT measures overcome the issue of ordinal scores, a questionnaire may still be time consuming to complete, especially if administered in paper-and-pencil format, due to the number of questions required to obtain robust measures (Bond & Fox, 2013; Andrich, 1988; Wright, 1977).

Before a CAT program is implemented, it is recommended that simulations be undertaken to evaluate testing parameters prior to live testing to ensure that the CATs will function optimally with the calibrated item bank. Three main types of simulations can be done, namely Post-Hoc, Hybrid and Monte-Carlo. A Post-Hoc simulation requires an existing item response matrix of real test takers for which item parameters have been estimated. The simulation uses the item responses to simulate how that bank would function if the items (with known difficulties) had been administered as a CAT. Such simulations can also be used to explore by how much the test length of a conventional test can be reduced by administering a test as a CAT.

One significant problem with Post-Hoc simulations is to have a data set in which all test takers have responded to all items in the bank. Banks are usually developed from different combinations of items included in different tests and all items in such banks are very seldom responded to by all test takers. To overcome

the limitation of a sparse item response matrix, Hybrid simulations can be used. A Hybrid simulation uses a calibrated item bank to estimate abilities for test takers with the available item responses which are then used to impute responses to un-administered items to generate a complete data matrix for implementation of Post-Hoc simulation. Monte-Carlo simulations can be used if little or no data is available. Monte-Carlo simulation is a computer simulation that allows the evaluation of various combinations of CAT options by using hypothetical model-generated test takers.

Simulations can get complex when a large number of conditions and/or multiple criteria are analyzed. Furthermore, one set of randomly generated data can be idiosyncratic in especially Monte-Carlo simulations. A number of replications of each condition is therefore recommended. If a testing program has quadrature points that are used across different item pools, multiple simulations at given abilities (thetas) can be used.

It is important to specify the aim(s) of the simulations beforehand to guide the decisions to be made about the number of items (whether fixed or variable), the termination rule, flexibility in content constraints, the maximum acceptable standard error of measurement, and so on.

In this study the efficiency of CATs is explored through simulations for a multiple-choice test and a five-category item questionnaire.

Case study 1: Multiple-choice question test

The purpose of the first case study was to investigate whether CATs can be used to reduce the test length of 60-item fixed-length tests without compromising measurement precision. An item bank with 260 four-choice dichotomously scored items from which 60-item fixed-length tests are compiled was available for the study. Following Rasch calibration and linking, the item difficulties were fixed with a minimum item difficulty of -4.857 logits, a maximum item difficulty of 3.143 logits and a mean item difficulty of -0.377 logits. The tests have been administered to test takers and normal distributions of ability estimates were generally observed. These tests had classical (Kuder-Richardson 20) reliabilities (Crocker & Algina, 1986) in the order of 0.70.

Method

Monte-Carlo simulations were based on the bank of 260 items with known item difficulties and 1 000 simulated test takers. Since practical considerations had to be borne in mind, the parameters in the simulations were gradually changed, taking the measurement error (SEM) into consideration. The intention of investigating the range of results was to allow for a viable balance between precision and the number of items required. Although the highest precision is desirable, it may not be a significant improvement in the number of items required when compared to the 60-item fixed length tests.

Abilities of the hypothetical test takers were initially assumed to be normally distributed ($N \sim (0, 1)$) in the range [-3; 3] logits. Alpha and beta values were used to control the beta distribution to mimic the actual distribution of abilities observed for the 60-item fixed length tests as closely as possible with the model constant set at 1.0 as the pure logistic model. The initial ability estimate was set in the range [-1; 1] logits. Ability (theta) was set to be estimated by maximum likelihood as implemented in the CAT algorithms and subsequent items were selected by maximum (Fisher) information at the current ability estimate.

To minimize idiosyncrasies in the simulations, different random seeding was used in a number of replications of the same and different requirements, no constraints were specified and uni-dimensionality was assumed.

In the first eight simulations the measurement precision was increased in three sets of simulations from a SEM of 0.40 through 0.35 to 0.30 assuming a normal distribution of abilities between -3 and 3 logits to determine the minimum and maximum number of items that would be required to achieve the precision. In the first three simulations no restriction was placed on the number of items to achieve the three specified SEMs; in simulations 4 to 6 the minimum and maximum number of items were specified as 15 and 35 respectively and based on these results the minimum and maximum number of items was increased to 20 and 40 in simulations 6 to 8. The purpose of running these sets was to find out how many items are required to achieve the three specified SEMs from the given pool of items. Table 1 summarizes the termination options for these simulations.

Table 1. Termination options for the first eight simulations

Simulation	SEM	Min # items	Max # items
1	≤ 0.40	-	-
2	≤ 0.35	-	-
3	≤ 0.30	-	-
4	≤ 0.40	15	35
5	≤ 0.35	15	35
6	≤ 0.40	20	40
7	≤ 0.35	20	40
8	≤ 0.30	20	40

This information is shown graphically in Figure 1.

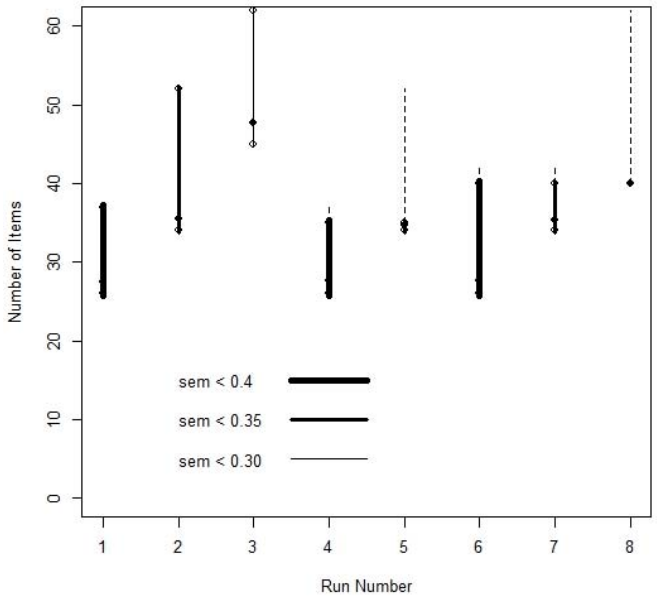

Figure 1. Plot of the range of items required to achieve three SEMs for eight simulation runs.

Although the 60-item fixed length tests yielded normally distributed abilities, the robustness of the intended CATs was investigated. In the ninth and tenth simulations rectangular ability distributions between -3 and 3 logits and between -2 and 2 logits respectively were generated. For the ninth simulation the SEM was retained at a maximum of 0.35 within 20 to 40 items and an ability range of -3 to 3 logits. However, the Alpha and Beta values were both changed to 1 to create a uniform distribution rather than a normal distribution. This was repeated in the tenth simulation except for changing the ability range to [-2; 2] logits in which the majority of ability estimates are located.

To further explore the robustness of the CATs, the SEM was retained at a maximum of 0.35 within 20 to 40 items and an ability range of -3 to 3 logits in the 11th and the 12th simulations. However, the Alpha and Beta values were altered to simulate skewed distributions – positively in simulation 11 and negatively in simulation 12.

Although fixed item difficulties were used in the simulations and no constraints were imposed, ability measures for hypothetical test takers were simulated using Monte-Carlo and therefore idiosyncratic information can be contained in the simulations. Some replications for the same conditions were thus deemed necessary. Based on the initial results, it was decided to focus on an SEM ≤ 0.40 in normally distributed simulations with abilities in the range [-3; 3] logits. A series of five simulations was run. For each simulation a different ability distribution was generated.

Results

The results of the first set of eight simulations is shown in Table 2. In the first three simulations the number of items to be administered was unbounded. To achieve measures with SEMs ≤ 0.401 it was found that a maximum of 37 item was required for all test takers. This level of precision could be achieved with 30 or less items for 95.4% of the test takers. Simulation two required higher precision at SEM ≤ 0.352 which was achieved with 52 or less items for all test takers. Note that with 40 items this precision can be achieved for 97.1% of the test takers and with 36 items for 80% of the test takers. Simulation 3 further increased precision to SEM ≤ 0.303 which was achieved with 62 items for all test takers. Note that this precision can be achieved with 50 items for 92.9% of the test takers and with 48 items for 78.6% of the test takers.

In summary, with a maximum of 37 items reliabilities of measures for all test takers can exceed 0.8 and with a maximum of 62 items reliabilities above 0.9 can be expected for all test takers. The fixed-length tests comprise of 60 items each. The simulations suggest that

[1] This can be interpreted as a classical reliability in the order of 0.84.

[2] Classical reliability in the order of 0.87.

[3] Classical reliability in the order of 0.90.

Table 2. Summary results of the first eight simulations

Simulation	1	2	3	4	5	6	7	8
Mean # items admin	27.54	35.52	47.63	27.58	34.67	27.61	35.34	40
Min # items admin	26	34	45	26	34	26	34	40
Max # items admin	37	52	62	35	35	40	40	40
% with SE ≤ 0.400	100	-	-	99.5	-	99.6	-	-
% with SE ≤ 0.350	-	100	-	-	33.1	-	96.7	-
% with SE ≤ 0.300	-	-	100	-	-	-	-	0

with 62 items SEMs ≤ 0.400 can be achieved for all test takers. For 30 items this level of precision can be achieved for 95.4% of the test takers.

In the fourth and fifth simulations the number of items were bounded to a minimum of 15 and a maximum of 354 items. The minimum number of items had no effect. In accordance with simulation one, SEMs ≤ 0.40 could be achieved with a maximum of 35 items for 99.5% of the test takers. Simulation five increased precision to SEMs ≤ 0.35. In accordance with simulation two, this precision could only be achieved for 33.1% of the test takers with 35 items. Increasing precision from a SEM of 0.40 to 0.35 thus had a significant effect if a maximum of 35 items is specified.

Simulations six, seven and eight retained the precisions specified in simulations one, two and three, but the number of items were bounded to a minimum of 20 and a maximum of 40. Again, the minimum number of items had no effect. The results for simulation six, as expected, were almost identical to the results of simulation four. If simulations five and seven are compared it is observed that increasing the maximum number of items from 35 to 40 makes a very significant difference – almost tripling the number of test takers with SEMs at or below the specified 0.35.

Simulation eight explored higher precision at SEMs ≤ 0.30 with an upper limit of 40 items. Simulation three indicated that a minimum of 45 items is required to achieve this implying that this precision is not possible for 40 or less items as the maximum number of items is reached before the precision is achieved.

The ninth simulation suggested that a SEM ≤ 0.35 can be achieved for 83.9% of the test takers and the tenth simulation found that this could be achieved for 95.7% of the test takers. For the normal distribution, see simulation seven, it was found that this could be achieved for 96.7% of the test takers. The results are thus very comparable if test takers with "extreme" (outside the ability range [-2; 2] logits) is excluded – which makes the two distributions much more comparable.

For the positively skewed distribution of simulation 11, it was found that SEM ≤ 0.35 can be achieved for 96.1% of the test takers and for the negatively skewed distribution of simulation 12 for 88.7% of the test takers. The results of simulations 7, 9, 10, 11 and 12 were compared since the parameters were the same except for the shape of the distribution and the range.

Table 3. Comparing the results of simulations 7, 9, 10, 11 and 12

Simulation	7	9	10	11	12
Mean # items admin	35.34	36.39	35.58	35.60	36.02
Min # items admin	34	34	34	34	34
Max # items admin	40	40	40	40	40
% with SE ≤ 0.35	96.7	83.9	95.7	96.1	88.7

A normal distribution (simulation 7) in the ability range [-3; 3] logits yields results equivalent to a rectangular distribution (simulation 10) in the ability range [-2; 2] logits. If the rectangular distribution's ability range is widened (simulation 9) to [-3; 3] logits a SEM ≤ 0.35 is achieved for more than 10% less test takers, i.e. the test takers at the extremes. A positively skewed distribution of abilities (simulation 11) had little effect on the result whilst a negatively skewed distribution (simulation 12) yielded marginally poorer results.

[4] The rationale for the lower limit is to investigate rapid convergence and non-convergence for the upper limit.

Table 4 summarizes the results obtained from the series of five replicated simulations for SEM ≤ 0.40 in normally distributed simulations with abilities in the range [-3; 3] logits. For each simulation a different ability distribution was generated.

This information is shown graphically in Figure 2.

The results from Table 4 suggest that the parameters are robust and stable and the mean values indicate that SEM ≤ 0.40 can be achieved with 37 or less items for all test takers. This is in the order of classical reliabilities around 0.84. Simulations 9 and 10 clearly demonstrated that much less items was needed if the ability range [-2; 2] logits instead of [-3; 3] logits is used. Simulations 11 and 12 suggested that some skewed distributions had little effect on the results.

A further simulation was run under the same conditions with SEM ≤ 0.50 (approximately corresponding to a classical reliability of 0.75). To achieve this, an average of 18.16 items was required with a minimum of 17 items and a maximum of 28 items. For 94.5% of the test takers this precision could be achieved with 20 items.

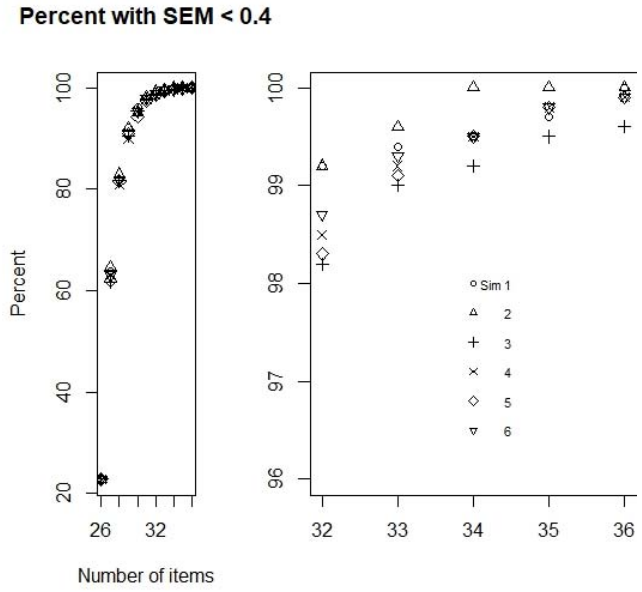

Figure 2. Plot of the number of items required against the percentage of test takers achieving SEM < 0.40.

Case study 2: Rating scale

The purpose of the second case study was to investigate to what extent CATs can be used to reduce the test length of a questionnaire consisting of 84 rating scale Likert-type questions. Test takers have to select one of five options "*Not at all*"; "*A little*"; "*Quite a bit*"; "*A lot*" and "*Extremely*" to statements.

Table 4. Five simulations for SEM ≤ 0.40

Simulation	1	2	3	4	5	Mean
Mean ability	0.003	-0.011	0.028	0.030	0.000	0.010
SD ability	0.893	0.904	0.943	0.928	0.873	0.908
Min ability	-2.381	-2.422	-2.605	-2.323	-2.495	-2.445
Max ability	2.421	2.218	2.526	2.587	2.524	2.455
% achieved	100	100	99.80	99.9	100	99.94
Mean # items admin	27.47	27.46	27.59	27.53	27.58	27.53
Min # items admin	26	26	26	26	26	26
Max # items admin	36	34	40	40	39	37.8
% after 36 items	100	100	99.6	99.9	99.9	99.9
% after 35 items	99.7	100	99.5	99.8	99.8	99.8
% after 34 items	99.5	100	99.2	99.5	99.5	99.5
% after 33 items	99.4	99.6	99.0	99.2	99.1	99.3
% after 32 items	99.2	99.2	98.2	98.5	98.3	98.7
% after 31 items	98.3	98.1	97.4	97.6	97.1	97.7
% after 30 items	96.2	95.2	95.0	95.2	94.1	95.1
% after 29 items	92.2	92.0	90.0	89.8	90.6	90.9
% after 28 items	81.9	82.8	81.1	80.9	81.2	81.6
% after 27 items	63.8	64.5	61.0	63.4	61.3	62.8
% after 26 items	22.8	22.4	22.0	23.2	21.7	22.4

Method

Following a Rasch rating scale calibration using data from paper-and-pencil administration of the questionnaire, difficulties for each category within each question were derived and located on a common scale. These difficulties were then used in a Monte-Carlo simulation to generate abilities for 1 000 hypothetical test takers with no constraints such as exposure or content. It was assumed that the pool of questions was uni-dimensional with locally independent questions. To minimize idiosyncrasies in the simulations, different random seeding was used in a number of replications of the same and different requirements. An initial simulation was based on normal (N ~ (0, 1)) distribution with abilities in the range [-3; 3] logits. No restrictions on the number of items were initially set and the precision in terms of the standard error (SE) was stepwise increased as SEM ≤ 0.50; SEM ≤ 0.40; SEM ≤ 0.35; SEM ≤ 0.30 and SEM ≤ 0.25 which can be approximately related to classical reliabilities of 0.75; 0.84; 0.88; 0.91 and 0.94 respectively. Robustness was explored through positively and negatively skewed distributions.

Results

The simulations suggested that only five questions were required to achieve a SEM ≤ 0.35 for 96% of the test takers in a normal distribution, and that between 11 and 18 questions were required in skewed distributions – instead of all 84 questions.

In order to investigate the predicted outcomes, CATs were administered to 113 people with ability estimates between -2.080 logits and 0.845 logits (mean of -1.182 logits) and the results are shown in Table 5.

Table 5. Results obtained for 113 people.

Minimum SEM	0.168 logits
Maximum SEM	0.573 logits
Mean SEM	0.329 logits

Discussion and Conclusions

In the first case study the viability of administering CATs was investigated for compiling tests from an item bank of 260 MCQs with known difficulties to improve on the 60-item fixed tests with reliabilities around 0.7.

Using the item difficulties, Monte-Carlo simulations were used to generate abilities for cohorts of 1 000 hypothetical test takers. The initial simulations were based on N ~ (0, 1) distributions with abilities in the interval [-3; 3]. No restrictions on the number of items were set and the precision was stepwise reduced as SEM ≤ 0.40; SEM ≤ 0.35 and SEM ≤ 0.30 which can be approximately related to classical reliabilities of 0.84; 0.87 and 0.91 respectively. It was found that these precisions could be achieved for all test takers with a maximum of 37; 52 and 62 items in each case. On average 27.52; 35.52 and 47.63 items were needed in each scenario.

In the second set of simulations these precisions were accompanied by restrictions to the number of items. A minimum of 15 and a maximum of 35 items was set. It was found that the minimum was not applicable since at least 26 items were needed to achieve the less precise measures, i.e. SEM ≤ 0.40. It was found that for SEM ≤ 0.40 the upper limit wasn't necessary since all test takers could be assessed with this precision. However, increasing the precision to SEM ≤ 0.35 had a significant impact – only around 33% of the test takers achieved this with 35 or less items. It was not meaningful to further increase the precision to SEM ≤ 0.30.

A third set of simulations repeated the second set but increasing the lower limit of the number of items to 20 and the upper limit to 40. Again the lower limit had no impact. The SEM ≤ 0.40 was achieved for all test takers and the SEM ≤ 0.35 by almost all test takers. However, changing SEM ≤ 0.30 had a significant impact and this precision could not be achieved by any test taker in 40 or less items. The third simulation where no limits on the number of items was specified indicated that a minimum of 45 items is required.

A fourth set of simulations investigated the shape of the distribution of ability measures. Whereas the previous simulations were based on normal distributions, uniform distributions were simulated. For the ability interval [-2; 2] logits the difference between the results obtained for the normal and the uniform distributions were negligible. However, if the same ability distribution [-3; 3] logits is used, the difference in results is significant with a difference of more than 10%. This can be interpreted that the bank does not have sufficient items at the lower and the upper difficulties to obtain measures at the specified precisions for test takers with abilities below -2 and above 2 logits for the uniform distribution.

The fifth set of simulations investigated skewed distributions, one negative and one positive. It was found that the former yielded similar results than the normal distribution whilst the latter had slightly worse results. This is due to more items at the bottom end of the difficulty continuum than at the top.

Settling on SEM ≤ 0.40 a series of simulations was done for N ~ (0, 1) in the ability interval [-3; 3] logits to inspect the robustness of the results. It was found that the results were stable and it can be concluded that at least 26 items and at most 37 items are needed to achieve this precision. Furthermore, this precision can be achieved for around 95% of the test takers with 30 or less items. In other words with half the number of items in the fixed-length tests reliabilities in the order of 0.84 can be achieved. A final simulation was done with SEM ≤ 0.50, i.e. reliability in the order of 0.75. It was found that an average of 18 items (a minimum of 17 and a maximum of 27 items) is required. This precision can be achieved for about 94% test takers with 20 items, i.e. a third of the number of items included in the 60-item fixed tests.

For the questionnaire the simulations suggested that only five questions were required to obtain SEMs ≤ 0.35 in a normal distribution and between 11 and 18 questions were required to achieve this precision in skewed distributions. Administration of seven questions yielded a mean SEM of 0.329 with the highest SEM at 0.573. Some 61 of the 113 people (54%) had SEMs less than 0.35 and 98 of the 113 people (86.7%) had SEMs less than 0.40.

Participants took 364.62 seconds (6.08 minutes) on average to respond to the seven questions. Some 78 of the 113 people (69.03%) took less than 7 minutes (one minute per question) and 104 of the 113 people (92.04%) took less than 14 minutes (2 minutes per question) to respond to the seven questions. If all 84 questions were administered, it would have taken the participants 12 times longer to complete the questionnaire.

The results from these two case studies firstly showed that simulations can predict what can be expected for CAT administrations and secondly that CATs significantly increase the efficiency of assessment without compromising measurement precision.

References

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

Barnard, J.J. (2015). Implementing a CAT: The AMC experience. *Journal of Computerized Adaptive Testing.* 3, 1-12.

Bond, T.G. & Fox, C.M. (2013). Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed). New Jersey: Lawrence Erlbaum Associates.

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

Van der Linden, W.J. & Glas, C.A.W. (Eds.). (2003). Computerized adaptive testing: Theory and practice. Dordrecht: Kluwer.

Wagner-Menghin, M.M. & Masters, G.N. (2013). Adaptive testing for psychological assessment: How many items are enough to run an adaptive testing algorithm? *Journal of Applied Measurement.* 14(2), 1-12.

Wang, T. & Vispoel, W.P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement.* 35, 109-135.

Weiss, D.J. (2011) Better data from better measurements using computerized adaptive testing. *Applied Psychological Measurement.* 2(1), 1-23

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement* 6, 473-492.

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

Citation:

Barnard, John J. (2018). From Simulation to Implementation: Two CAT Case Studies. *Practical Assessment, Research & Evaluation*, 23(14). Available online: http://pareonline.net/getvn.asp?v=23&n=14

Corresponding Author

John J. Barnard
Excel Psychological and Educational Consultancy
P O Box 3147
Doncaster East VIC 3109
Australia
email: john [at] epecat.com