**Pythagoras - Journal of the Association for Mathematics Education of South Africa**
ISSN: (Online) 2223-7895, (Print) 1012-2346



# Investigating the treatment of missing data in an Olympiad-type test – the case of the selection validity in the South African Mathematics Olympiad

**Authors:**
Caroline Long[1]
Johann Engelbrecht[2]
Vanessa Scherman[3]
Tim Dunne[4]

**Affiliations:**
[1]Department of Childhood Education, University of Johannesburg, South Africa

[2]Department of Mathematics, Science and Technology Education, University of Pretoria, South Africa

[3]Department of Psychology of Education, University of South Africa, South Africa

[4]Department of Statistical Sciences, University of Cape Town, South Africa

**Corresponding author:**
Vanessa Scherman,
scherv@unisa.ac.za

**Dates:**
Received: 02 Mar. 2016
Accepted: 15 Aug. 2016
Published: 31 Oct. 2016

**How to cite this article:**
Long, C., Engelbrecht, J., Scherman, V. & Dunne, T. (2016). Investigating the treatment of missing data in an Olympiad-type test – the case of the selection validity in the South African Mathematics Olympiad. *Pythagoras*, *37*(1), a333. http://dx.doi.org/10.4102/pythagoras.v37i1.333





The purpose of the South African Mathematics Olympiad is to generate interest in mathematics and to identify the most talented mathematical minds. Our focus is on how the handling of missing data affects the selection of the 'best' contestants. Two approaches handling missing data, applying the Rasch model, are described. The issue of guessing is investigated through a tailored analysis. We present two microanalyses to illustate how missing data may impact selection; the first investigates groups of contestants that may miss selection under particular conditions; the second focuses on two contestants each of whom answer 14 items correctly. This comparison raises questions about the proportion of correct to incorrect answers. Recommendations are made for future scoring of the test, which include reconsideration of negative marking and weighting as well as considering the inclusion of 150 or 200 contestants as opposed to 100 contestants for participation in the final round.

## Introduction

Mathematics competitions globally have grown into an immense vibrant network that engage millions of students and teachers, contributing significantly to the development and maintenance of mathematical knowledge and the educational process. However, performance in mathematics competitions does not always correlate with classroom performance (Ridge & Renzulli, 1981). In fact, Kenderov (2006) sees competitions as providing a tool to identify and develop students with higher abilities and talent who do not experience any challenge in the standard curriculum. The outcome of this curtailed curriculum experience in many a classroom is that the mathematical abilities and talent of students with great potential then remain undiscovered and undeveloped.

While they play an important role, competitions are not unconditionally promoted. One critique of competitions is that they provide unnecessary pressure, stress and feelings of failure from excessive competitiveness (Davis, Rimm & Siegle, 2011). Kenderov (2006) supports this view and argues that although students who perform well in competitions often become good mathematics researchers, many highly creative students do not function well under time pressure. He further states:

> What matters in science is rarely the speed of solving difficult problems posed by other people. More often, what matters is the ability to formulate questions and pose problems, to generate, evaluate, and reject conjectures, to come up with new and nonstandard ideas. (Kenderov, 2006, p. 1592)

The South African Mathematics Olympiad (SAMO), organised by the South African Mathematics Foundation (SAMF), is the premier mathematics Olympiad in the country and an important event in the school calendar. Participation has grown from just over 5000 contestants in the first event in 1966 to about 82 000 contestants who participated in the SAMO 2015. Objectives of the SAMO are to generate enthusiasm and interest in the subject, to enrich the study of mathematics, to promote mathematical problem-solving proficiency, to equip contestants for university level mathematical thinking and, in addition, to identify and inform selection of the finest young mathematical minds for international competition.

The reported benefits and critique of competitions apply in part to the SAMO. However, ongoing evaluations have over the years introduced changes to ensure the furthering of the main aims of the SAMO competition. This research study forms part of the evaluation in the interest of furthering the central objectives stated previously.

**Note:** Our respected and loved colleague, Tim Dunne, sadly passed away in a car accident after this article had been submitted for publication. We acknowledge his valuable contribution not only to the article but also to the field of Rasch measurement theory.

This study focuses on the second round results of the SAMO 2012 for the junior division (Grade 8 and Grade 9). We investigate specific aspects of the second round test for the purpose of providing information on the fitness for purpose of the test and for the future refining of the processes of contestant selection for the third and final round of the competition.

The primary question guiding this study is:

> *How do the testing procedures and processes, including administration, marking and analysis processes, support the selection of the most deserving 100 contestants for participation in the third round; in other words: are there contestants excluded from the third round who should reasonably be included?*

Subsidiary questions, supporting the primary question, include:

> *To what extent are the psychometric properties robust enough to claim measurement of mathematical excellence?*
>
> *How does the treatment of missing data affect the selection outcomes?*
>
> *How do various test design features, administration, scoring and analysis procedures affect the outcomes?*

## Design of the testing programme

At the design stage the construction and the selection of items is the task of *the Olympiad committee* comprising master teachers of mathematics, past Olympiad winners and mathematicians from tertiary institutions.

Items in the SAMO test are ranked from 1 to 20 according to the level of difficulty as judged by the item writing team. Items 1 to 5 are considered to be easy or accessible to most contestants; items 6 to 15 are considered to be moderately difficult and items 16 to 20 are expected to be most difficult. Each band of items is scored differently. The standard scoring procedure in SAMO is depicted in Table 1.

The question here is whether the weighting of items (as in Table 1) skews the selection outcome to some extent, whether the a priori judgement of weighting is valid and whether this weighting makes a difference to the ranking of contestants. These questions are explored in the results section and suggestions are offered in the discussion.

One of the features of the test programme is 'negative' marking. In the administration phase of the SAMO 2012 test, the instructions to contestants state that incorrect answers will be penalised. The rationale for negative marking is as follows. For each multiple-choice item with five options, the probability of answering correctly through random guessing is 20%. Taking this logic further over all the items in the test means that any contestants simply guessing all the way through the test will obtain scores close to an average score 20%. The intended outcome of the negative marking is to ensure that the random guessers score on average zero on each item.

**TABLE 1:** Weighting and penalty procedure and random guess score.

| | Correct answer | Wrong answer | No answer |
|---|---|---|---|
| Items 1–5 | 4 | –1 | 0 |
| Items 6–15 | 5 | –1 | 0 |
| Items 16–20 | 6 | –1 | 0 |

The intended behavioural effect is to eliminate random guessing. By knowing there is a penalty for wrong answers the contestant, when confronted with an item for which they find no apparent correct answer, is induced to omit the answer rather than risk the negative penalty, as opposed to a minimal chance for a prospective positive score. In the case where a contestant is not completely unsure of the answer but thinks they may select a correct option, they may take the risk.

The question arises here whether, when analysing the data, the induced missing value should be allocated a zero score or should be taken as missing and what effects these two different treatments have on the rankings of contestants.

The time allocated for the first round paper of the SAMO is one hour for 20 multiple-choice questions. For the second round the time is increased to two hours to complete the 20 multiple-choice questions. The test designers and administrators are of the opinion that the time is adequate for the test to serve its purposes. One reported caveat is that some good mathematical thinkers may require extra time per item simply because they are excessively thorough. This comment is in line with Kenderov's (2006) call for acknowledging and encouraging original mathematical thinking. The third round paper is four hours for six extended format questions. This format shows a radical departure from the first two rounds.

## Missing data

Missing data is common in research and alludes to planned and desired information that is not available for examination and analysis (Tsikriktsis, 2005). Thus, explanations for the missing data may be difficult to deduce (Mallinckrodt et al., 2003). It is due to the very nature of the phenomenon that it cannot be adequately described (McKnight, McKnight, Sidani & Figueredo, 2007). There are several reasons why data could be missing: these conditions may relate to the participants, the study design and the interaction between the participants and the study design. The contestant could have missed an item, saved the item for later and run out of time or felt reluctant to answer the question (Sijtsma & Van der Ark, 2003). Missing data may be described in terms of three mechanisms of missingness: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Missing completely at random (MCAR) asserts a completely unsystematic pattern. The probability of a missing data element at any observation on any variable is unrelated to any of the data values intended in the data set, missing or observed. We may write an equation to state the marginal probability and the joint conditional probabilities:

$$Prob\ (x_k \text{ is missing}) = Prob\ (x_k \text{ is missing} \mid x_k \text{ and } x_{obs})$$

An example of the MCAR phenomenon occurs when a contestant accidentally skips an item. The accidental skipping of an item has a probability that is not related to proficiency in the test construct, for example, and can therefore be classified as MCAR.

Data values are said to be MAR when the propensity for an element in a data case to be missing may be related to the observed values for that case, but the propensity is not related to the values of its own unobserved (missing) variables. Here the word random may appear to have an unfamiliar connotation. Equivalently, if any two data cases share identical values on their commonly observed elements, the pair will presumably have the same statistical behaviour on the other observations, whether observed or not. The probability of a missing value may depend upon the value of the observed elements. For example, the missing data mechanism may be related to general language proficiency or may be related to speed and time available, rather than the difficulty of the item or proficiency of the contestant. The relationship between proficiency, speed and difficulty impacts on the performance in any timed test.

In both instances, MCAR and MAR, the response mechanism is termed ignorable. Thus the researcher can make a reasoned argument to ignore the unknown factors leading to the missing data and thus permit a simpler approach to the available data (Pigott, 2001).

When unobserved data are neither MCAR nor MAR, the data are termed to be missing not at random (MNAR). MNAR means that the data is missing for a specific reason (i.e. the unknown value of a variable that may become missing in a data case may affect the probability of the value becoming missing). This situation arises in the context of this Olympiad study, when the missing status for an unobserved (missing) data value is directly attributable to contestant proficiency or item difficulty. In the case of negative marking being a deterrent to guessing, MNAR may be the case. However, there is an argument to be made that overcautious contestants may forego the reasonable probability of answering correctly.

Missing data, of any kind, influences the interpretation of the results. In the case of this study, we posit that the missing data impacts somewhat on the item difficulty estimates and on the ranking of the contestants. It therefore has consequences for validity and reliability claims (McKnight et al., 2007). It remains one of the major challenges for analysis, due to the fact that information has been lost. Potential solutions to address missingness include regarding omitted items as not administered, or by contrast as incorrect, and allocating a zero score (Ludlow & O'Leary, 1999).

In this SAMO 2012 junior second round test, 92.8% of the total item-person data points (20 × 4141 = 82 820) were recorded, with some 7.2% of the data missing, which may be perceived as relatively inconsequential. Nevertheless, the approaches to handling of missing data account for slight differences in the estimated locations of items and greater corresponding shifts in estimated person locations, which in turn account for variation in the selections of the top 100 contestants (at the highest person locations).

In this article each of these approaches is explored with reference to specific items and the consequences or the differing effects on the selection of contestants are discussed.

# Methodology

Our data sources consisted of 4141 junior contestants (Grade 8 and Grade 9) who participated in the SAMO 2012 second round test comprising 20 multiple-choice items with five options, one of which was correct. In view of the primary research question, that is whether or not the testing procedures and processes support the selection of the most deserving 100 learners, with particular reference to the procedure for handling missing data, two parallel analyses were conducted on two versions of the dataset. In one analysis, zeroes were assigned for all missing data (ZM) and in the second all missing data were handled through the standard procedures of the RUMM 2030 software (MM). For both sets of data, a Rasch analysis was conducted and the set of statistical methods applied to provide information on the test as a whole and on the individual items. Parallel results are reported throughout the analysis.

## Rasch measurement theory

The application of the Rasch measurement theory (RMT) to item level data collected on the second round tests enabled us to answer questions concerning the robustness of the data, in terms of the psychometric properties required for measurement-like outcomes. The model is clearly explained elsewhere (Andrich, 1988, 2004; Dunne, Long, Craig & Venter, 2012; Rasch, 1960/1980; Wilson, 2005); some pertinent aspects are discussed here.

The *dichotomous model,* scoring correct responses as one and incorrect answers as zero, is operationalised in the RUMM2030 software features and applied in the early stages of the analysis to verify the functioning of the test as a whole and the coherent functioning of individual items.

The model assumes that the probability of a contestant answering any dichotomous item correctly is 'a logistic function of the relative distance between the item location and the [contestant] location on a common linear scale' (Tennant & Conaghan, 2007, p. 1359). The Rasch analysis aligns both item difficulty and person proficiency on the same scale, by assigning estimated locations whose difference governs probabilities of zero or one item scores.

For person ($v$) and item ($i$), the probability of a correct response is governed by:

$$P\{x_{vi}=1\,|\,\beta_v,\delta_i\}=\frac{e^{(\beta_v-\delta_i)}}{1+e^{(\beta_v-\delta_i)}} \qquad \text{[Eqn 1]}$$

The left-hand side of the equation is read as 'the probability of [person] $v$ being correct [scoring 1] on item $i$ given the [person's] ability, $\beta_v$, and the item's difficulty, $\delta_i$'. The right-hand side involves only an exponential transformation of the difference between person ability, $\beta_v$, and the item difficulty, $\delta_i$. The function of the denominator in Equation 1 is to constrain the sum of the (two) probabilities for any dichotomous item to 1 (Andrich, 2006, p. 63).

One consequence of the model (Equation 1) being applied independently to all the data of each candidate and all items is an expectation that the parameters for candidate ability and item difficulty are all on a common interval scale, called the logit scale. If the data collectively fit the complete model derived from Equation 1, this common scale permits very specific stochastic interpretations of the observed data that are highly desirable, including measurement-like interpretations of the functioning of the test and its constituent items.

# Analysis

In this analysis there are two analytic processes that we regard as somewhat distinct but related in that they both impact on the outcomes of the contest and influence which second round contestants will be selected for the third round.

The first process is the analysis of the theoretical coherence of the test and the test items: the second is the method of scoring. We report on the model and the software selected for the analysis, the summary statistics (means, standard deviations, chi-square statistics and associated exceedance probabilities and person separation index) and then the fit of items and persons to the model.

The second process involves the handling of missing data and the possibility of randomly guessed correct items within the observed data. Its purpose is to manage the fact that despite the attempts to dissuade guessing, there would inevitably be one or more chance correct multiple-choice question responses guessed by a contestant for whom the conditional probability of a correct response on any difficult item seems very low. We expect that a contestant will have a low probability of success on difficult items precisely because the contestant appears to perform at a low ability level on the test as a whole.

For clarity, we note that weaker contestants tend to have lower scores and we expect them to offer correct responses to easier items. We also expect them by definition to either leave difficult multiple-choice questions items blank or possibly choose randomly amongst some or all of the options on that item subset. The effect of correct guesses made by weaker contestants is twofold. Each correct guess increases the candidate's score, but simultaneously makes the item in question appear less difficult than expected because the observed frequency of correct responses is increased by correct random guesses. The above rationale underpins the process in a tailored analysis (Andrich, Marais & Humphry, 2012).

While the focus of the article is essentially on the treatment of missing values in the Olympiad test, the first analysis was to discuss aspects of the test and the testing process that impact on the theoretical coherence of the test and the testing process. These aspects include the expected versus empirical difficulty, the fit of the test as a whole, the targeting of the test and investigation of specific items. A second analysis investigates the handling of missing data.

The ranks of the item locations that represent the empirical difficulty levels as determined by the Rasch analysis were compared with the ranks of the expected difficulty levels as assigned by the designers of the tests.

## Test validity

For the test as a whole, the summary statistics, including item and person means and standard deviations, point to the appropriateness of the test. The fit of the data to the model, reflected in the chi-square statistics, and the test reliability as reflected in the person separation index (PSI) provide evidence of the robustness of the test. The PSI, specific to latent trait models such as the Rasch model, contrasts the variance among the ability estimates of persons tested in the data relative to the error variance within each person (Andrich, 1982). The index provides a measure of internal consistency by providing an indicator of the separation of persons relative to the difficulty of the item. The equivalent in traditional test theory is the Kuder-Richardson 20, or Cronbach's alpha, which provides a measure of the internal consistency of the items, rather than a measure of person consistency relative to items (Andrich, 1982).

For measurement in the psychosocial sciences, as with the physical sciences, appropriate targeting of an instrument as a whole implies that maximum information can be expected in the functioning of the items over the persons who are to be assessed. Targeting is an implicit goal of the test designers, in seeking to select $k$ useful items as the instrument for their purpose (here using $k = 20$ items to select top-end mathematics talent on the basis of high total scores).

In the Rasch model, the item location mean is set by convention at zero as a reference value on the logit scale. The comparison of the item mean with the observed person location mean provides a summary post hoc indication of the appropriate targeting of the test for the observed contestant group. The closer the person mean to the item mean, zero, the more accurately measures of ability can be obtained.

For individual items, the *fit residual* is the standardised sum of all differences between observed and expected values summed over all persons. If a fit residual is over 2.5, or less than –2.5, the item is regarded as possibly misfitting. In the application of the Rasch model, the item fit is investigated and where anomalies are found these residuals are further investigated.

The sign of the item fit residuals relates to the notion of item discrimination. In the sense used here, discrimination is the rate of increase in the probability of a correct response to a specific item with respect to the underlying person ability

level. The Rasch model suggests expected frequencies of item scores for each ability level. If the observed pattern of item performance against ability levels exaggerates the frequency of zero scores among contestants of lower ability, and also exaggerates the frequency of correct scores for high ability contestants, the overall item-fit residual will be negative (smaller than expected), pointing to overly high discrimination. Overly high discrimination indicates that persons of high ability may be obtaining special advantage. This special advantage may indicate that a second dimension (possibly language proficiency) that is positively correlated with the intended construct gives learners of high ability an undue advantage (Masters, 1988). This possible second dimension is to be investigated.

The further analysis compared the two approaches to missing data; the first simply regarded the missing data as incorrect (by default) and entered zero (ZM). This approach implies a judgement was made that the reason the contestant did not answer the question was that the correct answer was unknown to the contestant, hence the data was MNAR and for a reason directly related to the construct knowledge being tested.

The second approach was to reserve judgement (assume that the data may be MAR, perhaps related to the time factor or a reluctance to offer a wrong answer) and rather to assign a value using maximum likelihood estimates of the parameters and of the corresponding expected values for the missing data elements. In this second approach the missing values are treated as simply absent (MM).

## Guessing and tailored analysis

It is inevitable, where the item difficulty far exceeds the proficiency of the candidate, that there may be guessing. Here a tailored analysis is conducted to adjust for possible guessing on each of the data sets, ZM and MM. This approach is based on the understanding that 'guessing is a function of the difficulty of the item relative to the proficiency of the person' (Andrich et al., 2012, p. 1). Andrich et al. (2012) propose a strategy to 'remove those responses most likely to be affected by guessing' (p. 3), making particular removals for each person based upon their pattern of scoring. These removed responses are then treated as missing responses.

A tailored analysis was conducted with the cut-off at 0.20 or one-fifth, based upon the frequency for a randomly selected option from the set of five options offered for each of the multiple-choice questions. This cut-off is applied to the estimated probability of correct response, obtained from the Rasch model. For each person all the difficult items with estimated probabilities of a correct response below 0.20 are eliminated, hence rendered missing, whether correct or incorrect.

## Identifying best contestants

The highest performing contestants in each of the two modes, zero for missing (ZM) and missing as missing (MM), and then in each of the further two tailored analyses, zero for missing tailored (ZMT) and missing as missing tailored (MMT) are ranked from highest to lowest. Contestants whose rankings on all four treatments fall exclusively within particular ranges are coded. For example, contestant P may be ranked 3rd in the ZM ranking, 6th in the ZMT ranking, 4th in the MM ranking and 7th in the MMT ranking. This contestant would be coded A, as all four rankings fall into the top 25. Contestant Q may be ranked 39th in ZM ranking, 47th in ZMT ranking, 13th in MM ranking and 22nd in MMT ranking. This contestant would be coded B as all four rankings fall within the top 50. The contestants with all four rankings in the top 100 would fall into category C. This process is continued for contestants' rankings falling within 150 and within 200 (see Table 2).

The count of 14 includes only the contestants who achieved rankings exclusively in the top 25, 25th inclusive, in all four statistical treatments. Likewise, the count 27 includes all contestants who achieved ranking in the top 50; it therefore includes the count of 14 who achieved rankings exclusively within the top 25 (see Table 2).

Table 3 shows the number of contestants who obtained the highest total score of 19, down to the total score of 12, in the vertical column. Along the top row are the number of missing

**TABLE 2:** Cumulative counts of contestants in rank groups on four analysis routines.

| A | B | C | D | E |
|---|---|---|---|---|
| All rankings in top 25 | All rankings in top 50 | All rankings in top 100 | All rankings in top 150 | All rankings in top 200 |
| 14 | 27 | 54 | 85 | 107 |

**TABLE 3:** Counts of contestants by total score and missing answers.

| Total score | Missing | | | | | | | | | Count | Cumulative |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | | |
| 19 | 4 | 1 | | | | | | | | 5 | 5 |
| 18 | 8 | 2 | 0 | | | | | | | 10 | 15 |
| 17 | 6 | 1 | 0 | 0 | | | | | | 7 | 22 |
| 16 | 13 | 3 | 2 | 1 | 0 | | | | | 19 | 41 |
| 15 | 14 | 9 | 3 | 1 | 3 | 0 | | | | 30 | 72 |
| 14 | 25 | 9 | 6 | 3 | 6 | 1 | 0 | | | 50 | 121 |
| 13 | 39 | 11 | 4 | 9 | 7 | 4 | 1 | 0 | | 75 | 196 |
| 12 | 98 | 16 | 5 | 11 | 9 | 11 | 6 | 1 | 0 | 157 | 353 |
| Cumulative | 208 | 51 | 20 | 25 | 25 | 16 | 7 | 1 | 0 | 353 | |

items for each group. For example, for the group whose total score was 16 (looking along the row), 13 contestants had no missing values, three contestants had one missing value, two contestants had two missing values and one contestant had three missing values. It appears from this analysis that negative marking did not dissuade the contestants from answering questions whose correct answer was not immediately obvious.

### Micro-studies

To further investigate the handling of missing data in the context of the SAMO we conduct two explanatory micro-studies for illustration purposes. In the first we compare three categories of contestants who would fall, according to this system, outside of the top 100. We identified three other categories, some 25 contestants who had 14 out of the 20 items correct and none incorrect (Category F), some 39 contestants who had all 13 items that they attempted correct (Category G) and some 26 contestants who had 12 out of 13 or 13 out of 14 (Category H). Contestants in these categories would fall out of the top 100, on one, two or three of the rankings but be included on others. One contestant from each of the categories F, G and H is described.

A second micro-study identifies two contestants each of whom obtained 14 correct answers; the first answered 14 correctly and six incorrectly and the second answered 14 correctly and omitted the other six items. This comparison raises questions about the effect of time on assessing mathematics ability. It also raises questions about the proportion of correct to incorrect answers. *Does the contestant who answers everything they attempt correctly have greater potential than the contestant who answers almost a third of the test incorrectly?*

# Results

Throughout this investigation we conduct a comparison between the two data sets, the first case where the zeroes are given for missing information, denoted as ZM, and the second case where missing values are dealt with through maximum likelihood estimation, denoted as MM.

The results are presented in terms of both the test as a whole and the component items and on the performance of the contestants. The following analytic categories are discussed: *expected and empirical item difficulty*, *summary statistics* on the test as a whole, *targeting of the test*, *individual item statistics* and the *guessing factor*. Two additional categories that might have been reported, namely differential item functioning and local independence, which may have been important for checking the robustness of the data, have not been covered here.

The contestants are discussed in relation to their overall performance and in relation to the varied rankings based on the different analytic treatments, namely ZM and MM, and in pertinent cases the tailored equivalents, MMT and ZMT.

## Test design

The *notional (or expected) difficulty of items* from the perspective of the test designers was found to differ from the empirical difficulty outcomes. In 50% of the cases the items were at the expected levels (see Table 4).

The reasons for the differences are explored elsewhere (Engelbrecht & Mwambakana, 2016). Some of the reasons for the unexpected differences may include curriculum coverage, language issues, lack of exposure to the particular problem-solving strategy required for a specific problem or an element of surprise not obvious at the time of setting the paper.

Associated with the change in item difficulty order, there will be some difference in most contestants' scores when using the weighted scores, rather than using unweighted scores (as shown in Table 1).

## Summary statistics and targeting of the test

A comparison between the two versions of the data set is made, the first with zero scores for missing information (ZM) and the second case in which all missing values are dealt with through maximum likelihood estimation of person locations (MM).

The high total chi-square and the extremely low chi-square probability indicate that there is poor fit to the model (see Table 5). The PSI indicates rather moderate reliability with slightly improved reliability when the missing values are estimated through the standard procedure of the RUMM software. The moderate PSI indicates a limited spread of person locations along the scale. This limited spread may be an indication of a fairly homogeneous group taking the test, a consequence of common selection from a previous first round of testing serving as a screening mechanism.

We note here that the large number of data cases, over 4000, inevitably exacerbates the statistics for any misfit. By artificially reducing the size of the data set to a manageable number, the fit statistics were found to be more acceptable. Nevertheless, we use the fit statistics, prior to any statistical correction improvement (see Table 5).

**TABLE 4:** Comparison of intended and empirical difficulty, under ZM and MM methods.

| | Design of the test items ranked from less difficult to greater difficulty | Empirical ranking of items ranked from less difficult to greater difficulty | | Overlap between intended and empirical difficulty |
|---|---|---|---|---|
| | | ZM case | MM case | |
| Easy | 1→2→3→4→5 | 16→6→4→13→1 | 16→6→4→13→1 | 40% (2/5) |
| Moderate difficulty | 6→7→8→9→10 | 11→8→7→15→9 | 11→8→15→7→9 | 60% (6/10) |
| | 11→12→13→14→15 | 17→19→2→14→3 | 17→2→14→19→3 | |
| Difficult | 16→17→18→19→20 | 5→18→20→10→12 | 5→18→20→12→10 | 40% (2/5) |

**TABLE 5:** Test fit to the model.

| | ZM | MM |
|---|---|---|
| Total chi-square | 1742.53700 | 1900.93500 |
| Total degrees of freedom | 160.00000 | 178.00000 |
| Total chi-square probability | 0.00000 | 0.00000 |
| Person separation index (PSI) | 0.59346 | 0.62227 |

The statistics suggest that the test is moderately well targeted, with a person location mean of –0.804 (standard deviation 0.868) for the ZM case and a person location mean of –0.645 (standard deviation 0.942) for the MM case (see Table 6). The standard deviation in both cases indicates a spread of items that is acceptable. The optimal situation for obtaining maximum information is for the person mean to be aligned with the item mean. In the case of this test, the mean of –0.804 suggests the test is slightly on the difficult side for this set of contestants. The person mean in the MM case is located closer to the item mean, zero, and, given an argument for regarding missing data as missing rather than zero, could be regarded as the more accurate test statistic.

The person-item threshold distribution is depicted graphically in Figure 1. The spread of items is fairly good, with slightly better fit for the MM analysis. There is, however, a cohort of contestants for whom there are no items within their proficiency range. About 500 contestants, roughly an eighth of the complete set, have locations below the easiest item, Item 16. The implication here is that all of these contestants have a less than 50% chance of answering any specific item correctly. An explanation of this phenomenon is given in the discussion.

## Item difficulty

The person-item maps (see Figure 2, ZM, Figure 3, MM) are graphical pictures of both item difficulty and person proficiency aligned on the same scale. Items are depicted on the right-hand side of the graph. On the left-hand side, the estimates of learner proficiency are located. Item 10 and Item 12 are distinctly more difficult than other items. The next three items, 5, 18 and 20, form a second but less difficult cluster. Items 7, 8, 9, 11 and 15 form a cluster around the mean zero location. The remainder of the items spread from –0.2 to –1.8 on the logit scale, indicating relatively easier items. From an investigation of each cluster, it may emerge that particular problem-solving skills are required in addition to mastery of the topic area.

The estimated item difficulty locations differ across the missing data treatments. However, the group of the five most difficult items is common but in a slightly different order (see Table 7). Similarly, the five easiest items are in slightly different orders and likewise both middle quartiles involve common items but with internal order changes. The range of the MM analysis is slightly narrower, with the easiest item apparently slightly harder.

A tailored analysis, by adjusting for guessing, provides a better estimate of item difficulty and of item fit (Andrich et al., 2012). For this reason, we report for particular items the statistics provided by the tailored analysis MMT, as derived from MM. These results are presented in the three columns on the right of Table 7.

## Individual items

Item 10 (see Figure 4), one of the most difficult items, required an understanding of three concepts: the area of a circle, ratio and probability. A feature of this question was that the correct option was not the most frequent choice by even the top contestants. This selection can be explained by the first distractor being very seductive (most frequently chosen), while being incorrect. We report here on the statistics after applying a tailored analysis.

For each item discussed, there is a description of the item, together with its item characteristic curve and the multiple-choice distractor plot. For further explanation see Dunne et al. (2012).

Item 16, contrary to expectation, was found to be the easiest item empirically (see Figure 5). The easiness of the item can be explained by the strategy of some contestants to, rather than use mathematical techniques, simply extend the pattern to the 81st term.

## Item fit

The item fit residual is a statistic for assessing the extent to which an aggregate of the person-item residuals deviates from its expected value, zero. When fit residuals are greater than 2.5 or less than –2.5, we regard them as exceeding criterion levels for misfitting. In the initial data analysis six items were overdiscriminating (1, 3, 6, 14, 17, 19 with negative item residuals) and six items were underdiscriminating (2, 7, 10, 11, 15, 18). With further analysis (see the section on the tailored analysis), some of these items conformed to the model.

Item 2 (see Figure 6) initially exhibited large overdiscrimination (–7.883); however, after the tailored analysis, while still showing some overdiscrimination, the fit was better (–3.871). The percent topic is a complex construct as explained by Parker and Leinhardt (1995). The problem was that of ignoring the referent to which the percentage ratio is being applied. We might also posit here that additive reasoning is applied in this problem rather than multiplicative reasoning.

**TABLE 6:** Summary statistics.

| | ZM data ($N$ = 4141) | | | | MM data ($N$ = 4141) | | | |
|---|---|---|---|---|---|---|---|---|
| | Items | | Persons | | Items | | Persons | |
| | Location | Fit residual | Location | Fit residual | Location | Fit residual | Location | Fit residual |
| Mean | 0.0000 | –0.1947 | –0.8041 | –0.1476 | 0.0000 | –0.2580 | –0.6450 | –0.1362 |
| *SD* | 0.9440 | 4.4771 | 0.8675 | 0.9275 | 0.8940 | 4.6141 | 0.9416 | 0.8908 |

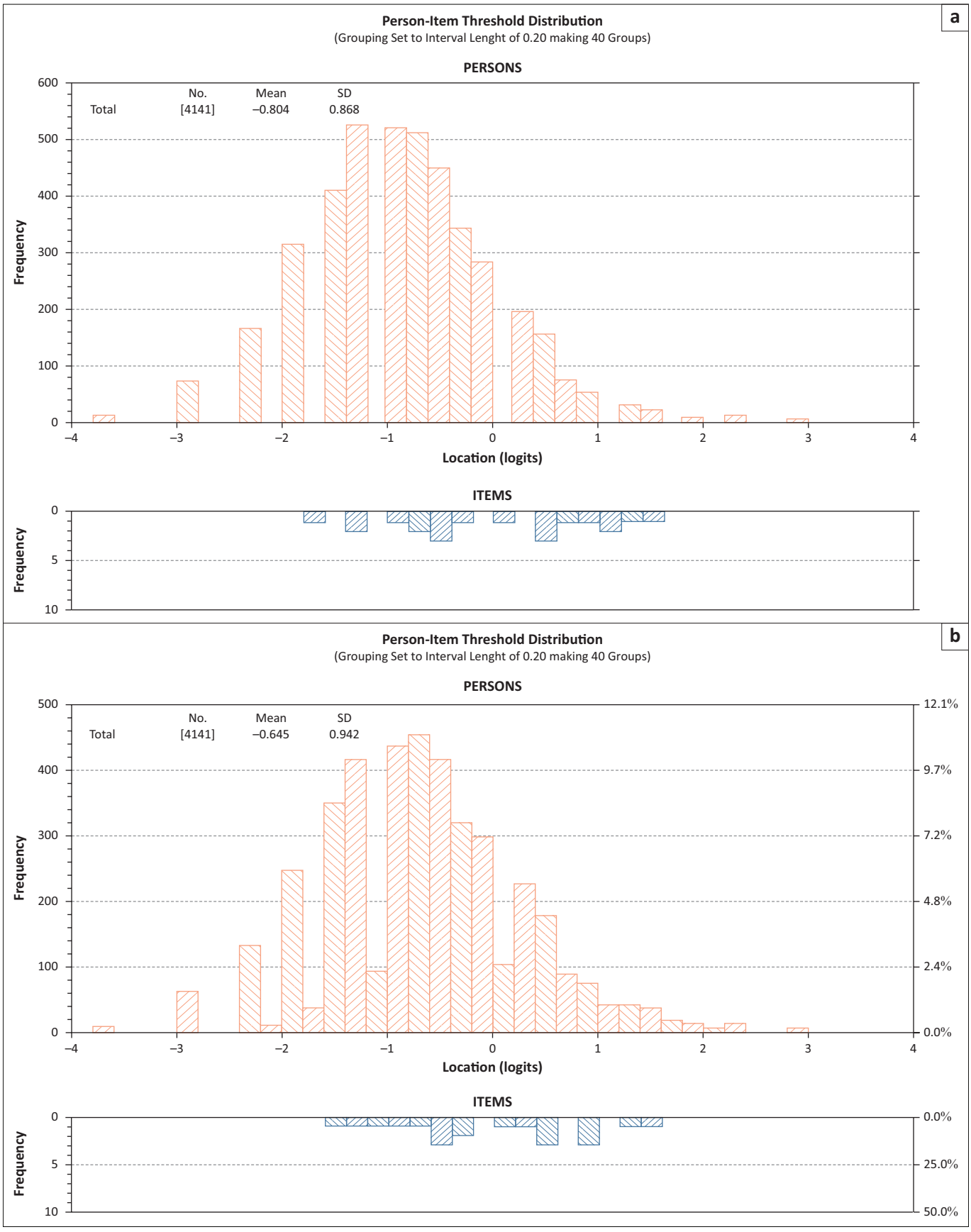


**FIGURE 1:** Person-item distribution thresholds for ZM (a) and MM (b) analyses.

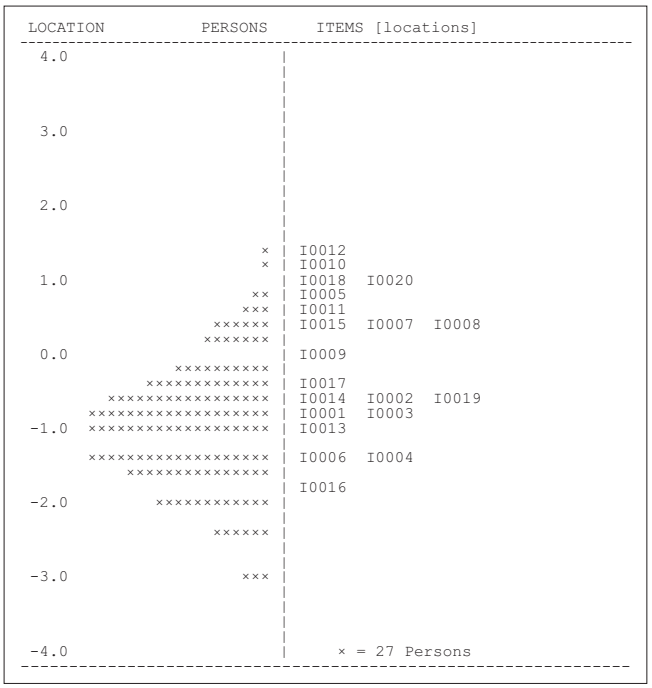


**FIGURE 2:** Person-item map for ZM data with mean item location at zero.

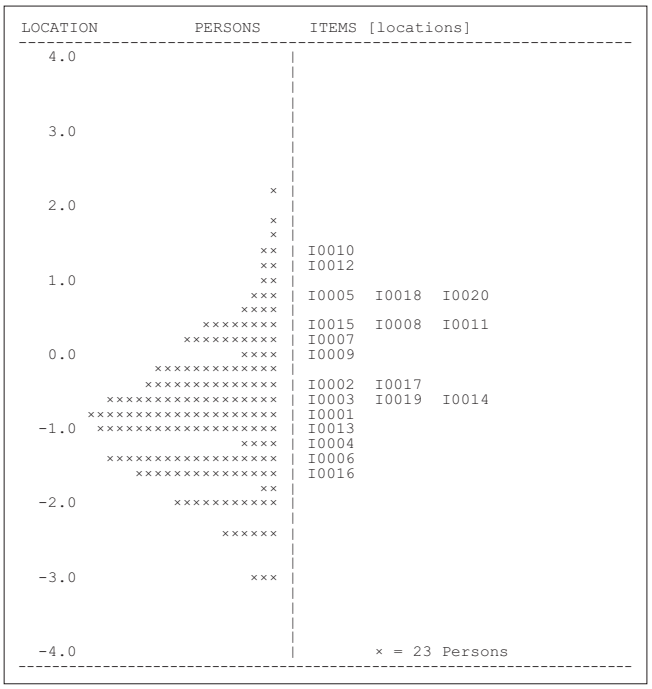


**FIGURE 3:** Person-item map for MM data with mean item location at zero.

**TABLE 7:** Item difficulty locations and standard error: ZM, MM and MMT methods.

| ZM | | | MM | | | MMT | | |
|---|---|---|---|---|---|---|---|---|
| **Item** | **Location** | ***SE*** | **Item** | **Location** | ***SE*** | **Item** | **Location** | ***SE*** |
| 12 | 1.401 | 0.049 | 10 | 1.405 | 0.049 | 10 | 2.118 | 0.089 |
| 10 | 1.381 | 0.049 | 12 | 1.390 | 0.050 | 18 | 1.562 | 0.073 |
| 20 | 1.186 | 0.046 | 20 | 0.994 | 0.048 | 20 | 1.538 | 0.078 |
| 18 | 1.173 | 0.046 | 18 | 0.986 | 0.048 | 5 | 1.216 | 0.062 |
| 5 | 0.977 | 0.043 | 5 | 0.923 | 0.045 | 12 | 1.074 | 0.078 |
| 11 | 0.604 | 0.039 | 11 | 0.555 | 0.040 | | Deleted | |
| 8 | 0.488 | 0.038 | 8 | 0.510 | 0.040 | 15 | 0.671 | 0.046 |
| 7 | 0.458 | 0.038 | 15 | 0.429 | 0.038 | 7 | 0.469 | 0.047 |
| 15 | 0.404 | 0.038 | 7 | 0.383 | 0.040 | 8 | 0.353 | 0.048 |
| 9 | 0.071 | 0.036 | 9 | 0.132 | 0.036 | 9 | 0.073 | 0.04 |
| 17 | –0.299 | 0.034 | 17 | –0.276 | 0.035 | 17 | –0.407 | 0.038 |
| 19 | –0.450 | 0.034 | 2 | –0.331 | 0.034 | 2 | –0.499 | 0.036 |
| 2 | –0.477 | 0.034 | 14 | –0.500 | 0.035 | 14 | –0.559 | 0.037 |
| 14 | –0.489 | 0.034 | 19 | –0.518 | 0.036 | 19 | –0.671 | 0.037 |
| 3 | –0.604 | 0.034 | 3 | –0.596 | 0.035 | 3 | –0.737 | 0.037 |
| 1 | –0.757 | 0.033 | 1 | –0.665 | 0.034 | 1 | –0.809 | 0.036 |
| 13 | –0.937 | 0.033 | 13 | –0.844 | 0.034 | 13 | –0.967 | 0.036 |
| 4 | –1.215 | 0.034 | 4 | –1.183 | 0.035 | 4 | –1.313 | 0.037 |
| 6 | –1.311 | 0.034 | 6 | –1.265 | 0.035 | 6 | –1.459 | 0.037 |
| 16 | –1.605 | 0.035 | 16 | –1.529 | 0.036 | 16 | –1.653 | 0.038 |

Item 6 (see Figure 7) on the initial analysis was highly overdiscriminating (–6.269). With the tailored analysis, the discrimination was reduced to –2.772. This item requires an application of algebra that is perhaps unfamiliar to many contestants. In addition, the requirement to explore the numbers in a number block may not have been previously encountered.

## Guessing and tailored analyses

As stated previously, additional procedures were administered that would provide a more accurate picture of the difficulty of the items, by firstly eliminating Item 11 (found to be faulty), and secondly by conducting a tailored analysis to adjust for guessing.

Table 8 shows a comparison of the summary statistics across the two approaches, the ZM, and the associated tailored analysis, ZMT, and the MM, and the associated MMT. Note that the person location means of the tailored analyses being closer to zero indicate that the test is better targeted in this approach. The reason for this better fit is that where an item is judged to have a low probability of being answered correctly by a specific person, that data point is omitted and

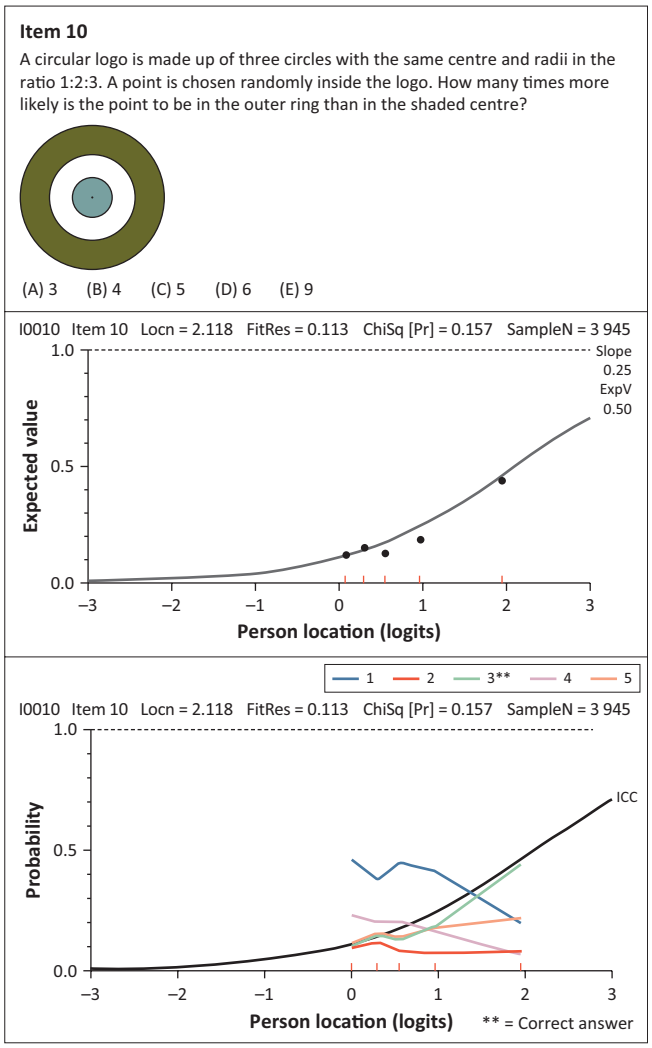


**FIGURE 4:** Item 10, with item characteristic curve, and multiple choice distractor plot.

regarded as missing. The reported person separation index improves in this situation (see Table 8).

With the application of the tailored analysis there was a better fit for some of the items, although some items emerged as worse fitting (see Table 9).

Further analysis, which explores possible reasons for some items improving with the tailored analysis and others not, as well as investigating local independence and the extent of guessing, is in process.

## Selection of contestants

The higher performing contestants, expected to have some chance of success in the final round, were clustered into five groups, as explained earlier, according to their performance in the four statistical treatments ZM, ZMT, MM and MMT. The number of contestants in each group is given in Table 10.

The practice applied in SAMO 2012 was to select the top 100 contestants according to their performance when the ZM

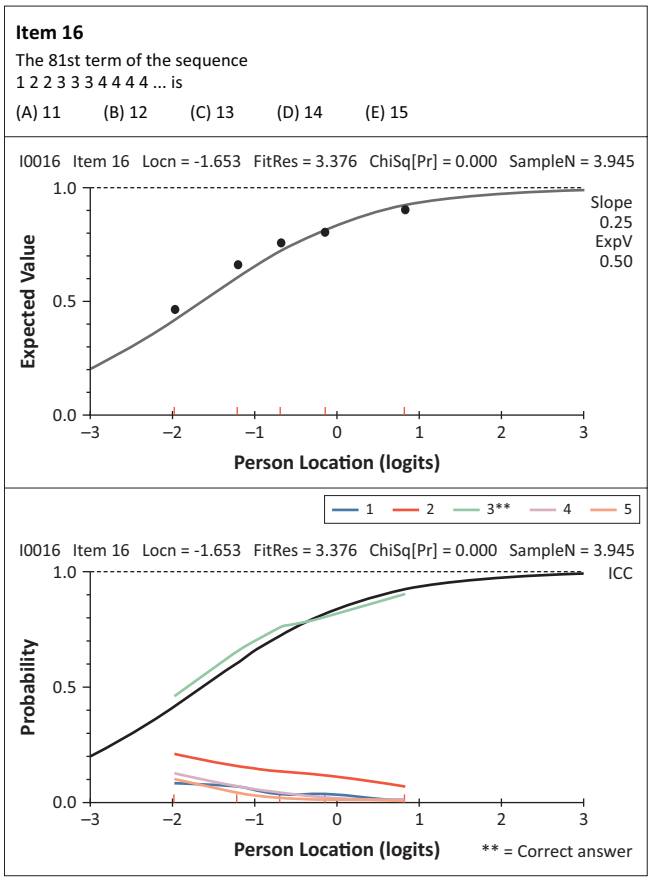


**FIGURE 5:** Item 16, with item characteristic curve, and multiple choice distractor plot.

approach is used. The analyses applied in this study seems to indicate that in order to select the most deserving 100 for the third round, it may be better to consider all contestants that fall into categories A to E.

Three other categories were also considered, comprising candidates who were not in any of the categories A to E. Thus we have category F, those contestants with a score of 14 out of 20, with no errors, category G, those contestants who scored 13 correct, with no errors, and category H, where contestants scored 11 or 12 correct with no errors, or obtained 11, 12 or 13, with only one error (Table 11). All of these contestants fall outside the top 100, but within the top 155 (see Table 11).

It is with the above reasoning that a closer micro-analysis was conducted on three contestants falling into the above three categories to use for illustrative purposes (see Table 12).

Fay scored 14/20 in the test. She qualifies for the top 100 if we use the ZM approach but does not when any of the other approaches is used. Grant only answered 11 items but answered no items incorrectly. Using the MM or MMT approach he qualifies for the top 100 easily but he does not when we use either ZM or ZMT. Henry attempted 13 items and answered 12 correctly. Again, when we apply either MM or MMT criteria he qualifies for the top 100 easily but not when we use ZM or ZMT.

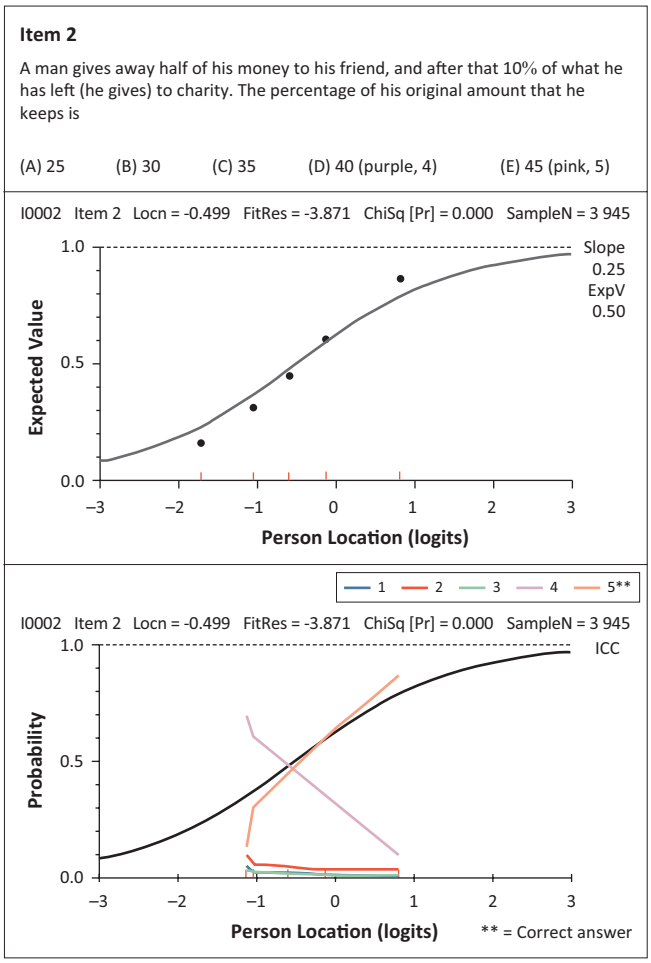


**FIGURE 6:** Item 2, with item characteristic curve, and multiple choice distractor plot.

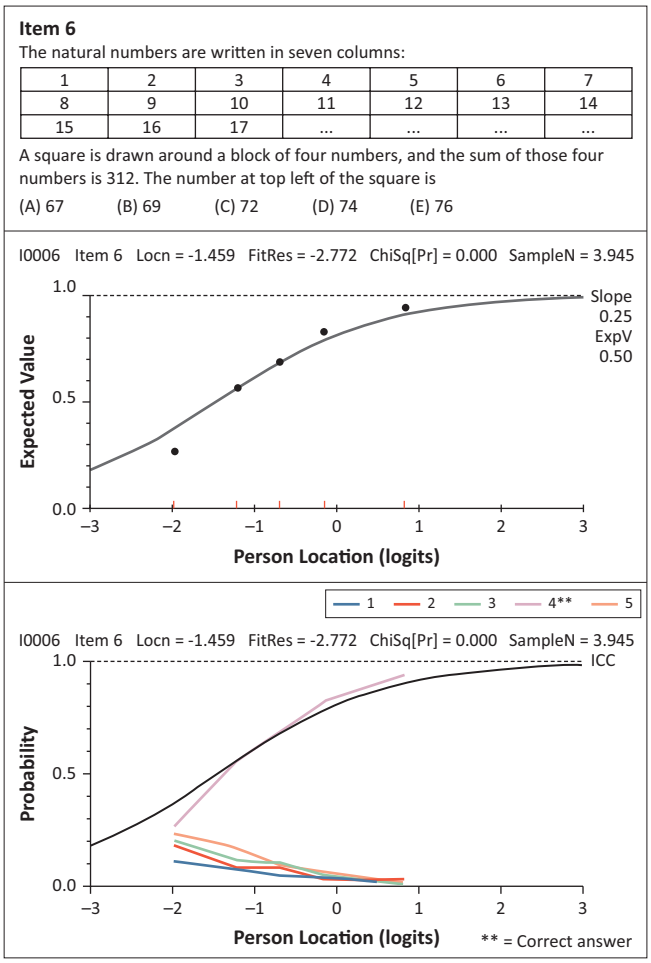


**FIGURE 7:** Item 6, with item characteristic curve, and multiple choice distractor plot.

**TABLE 8:** Comparison of statistics across tailored analyses.

| (*n* = 4141) | ZMT w/o Item 11 | | ZM Initial analysis | | MMT w/o Item 11 | | MM Initial analysis | |
|---|---|---|---|---|---|---|---|---|
| | Location | Fit residual | Location | Fit residual | Location | Fit residual | Location | Fit residual |
| Item mean | 0 | 1.4870 | 0 | –0.1947 | 0 | 1.1918 | 0 | –0.2580 |
| Item *SD* | 1.1952 | 3.3663 | 0.9440 | 4.4771 | 1.1210 | 3.5108 | 0.8940 | 4.6141 |
| Person mean | –0.8903 | 0.0747 | –0.8041 | –0.1476 | –0.7082 | 0.0422 | –0.6450 | –0.1362 |
| Person *SD* | 1.0196 | 0.8925 | 0.8675 | 0.9275 | 1.0788 | 0.8807 | 0.9416 | 0.8908 |
| PSI | 0.44890 | | 0.5935 | | 0.5251 | | 0.6223 | |

In a second micro-analysis, we compare the results of two contestants. Adri answered 14 items, all of them correctly. Her score for the test according to the system that was used, that is, taking the weighting and penalties into account, is 68%. Bets attempted all 20 items and 14 of her answers were correct. After weighting and penalties, her final mark was 64% for the test. In Table 13 we compare the two contestants. The items in Table 13 have been ordered according to the empirical difficulty (location) when analysed with missing values regarded as missing.

We see from Table 13 that although Adri and Bets answered the same number of items correctly (14), Adri obtained a final mark of 68% and Bets obtained a mark of 64%, because of the penalties for incorrect answers. It is also clear that most of Adri's missing answers were of high empirical difficulty. This pattern indicates that her missing answers could be classified as MNAR: she deliberately avoided answering the difficult questions of which she was uncertain. Bets lost most of her penalty marks in the more difficult items, where she possibly guessed but selected answers incorrectly. There is a high probability that Bets might have guessed Item 20 correctly. The fact that she answered Item 4 incorrectly is something of an anomaly.

## Discussion and recommendations

The results of the Rasch analysis show that the test is fairly robust and adequate for the purpose. Drawing from both professional judgement and from the outcomes of the analysis we conclude that the test is appropriate for engaging contestants' interest and enthusiasm; it is appropriately targeted with both easy items to encourage contestants at the lower end and challenging items at the

**TABLE 9:** Comparison of tailored analysis and MM case analysis from perspective of MM.

| MM fit residual order | | MM tailor fit residual order | | Direction of change |
|---|---|---|---|---|
| **Item** | **Fit Residual** | **Item** | **Fit Residual** | |
| 2 | –7.883 | 2 | –3.871 | better fit |
| 6 | –6.269 | 6 | –2.772 | better fit |
| 19 | –5.649 | 19 | –2.433 | better fit |
| 17 | –5.048 | 17 | –2.041 | better fit |
| 1 | –4.787 | 1 | –1.203 | better fit |
| 3 | –4.483 | 3 | –1.358 | better fit |
| 8 | –2.372 | 8 | 1.592 | better fit |
| 12 | –1.49 | 12 | –0.757 | better fit |
| 4 | –1.049 | 4 | 1.778 | change from slight overfit to slight underfit |
| 13 | –0.807 | 13 | 2.462 | change from slight overfit to slight underfit |
| **9** | **–0.31** | **9** | **4.423** | **change from slight overfit to marked underfit** |
| **16** | **1.079** | **16** | **3.376** | **worse fit** |
| 20 | 2.311 | 20 | 0.331 | better fit |
| 5 | 2.505 | 5 | –0.218 | better fit |
| **7** | **2.762** | **7** | **6.83** | **worse fit** |
| **14** | **2.994** | **14** | **5.167** | **worse fit** |
| 10 | 3.031 | 10 | 0.113 | better fit |
| 18 | 3.378 | 18 | 1.55 | better fit |
| **15** | **7.716** | **15** | **9.675** | **worse fit** |
| 11 | 9.209 | | | |

**TABLE 10:** Counts in contestant ranking categories within four analyses.

| Code | Category | Count | Cumulative |
|---|---|---|---|
| A | All 4 ranks in top 25 | 14 | 14 |
| B | All 4 ranks in top 50 | 13 | 27 |
| C | All 4 ranks in top 100 | 27 | 54 |
| D | All 4 ranks in top 150 | 31 | 85 |
| E | All 4 ranks in top 200 | 22 | 107 |

**TABLE 11:** Counts in further rank categories.

| Code | Category | Count | Cumulative |
|---|---|---|---|
| F | Score of 14 | 17 | 124 |
| G | Extreme scores 13/13 | 10 | 134 |
| H | 11, 12 or 11/12, 12/13 or 13/14 | 21 | 155 |

difficult end which will discriminate between the top contestants.

It appears that the time allowed was adequate and that most participants managed to reach the end of the test, though we do not know how the participants approached the test as a whole. The missing 7.2% of 4141×20 data points may have been due to time constraints. For the present purposes we acknowledge that the missing data may be impacted by time constraints and this fact should be considered in the handling of missing data.

Negative marking and weighting definitely impact the results. Negative marking increases the count of missing data elements. Weightings of item contribution to the final score needs serious reconsideration since the weighting is based on the anticipated rather than on the empirical location. We recommend no a priori weightings of item contribution to the final score. In the Rasch analysis even empirical weighting is redundant, because item location ensures contestants who can handle difficult items are credited for that ability. In the Rasch framework it is possible to construct open-ended items for which a higher maximum score than the value 1 can be allocated by a suitable memo.

Some aspects of the analysis may be informative for future test design. The cluster of contestants, about 500, for whom the test appears too difficult should perhaps have been omitted from the second round testing. A closer investigation of these 500 contestants indicate some anomalies in the system. A deviation from the normal selection procedure was made on account of providing opportunities to learners who were exceptional at their school but who did not meet the cut-off. The question arising is whether the positive benefits for participating outweigh possible knocks to confidence. SAMF currently have programmes in place for learners with potential to receive more tuition in problem-solving strategies earlier on in the Olympiad cycle, in an attempt to address lack of familiarity with problem-solving strategies in cohorts such as these.

The most important question that was addressed in this study is whether this testing process adequately supports the selection of the top 100 contestants who will go through to the third round of the Olympiad. This question does not have a simple answer and the only way of addressing this issue would be to monitor contestants' performance in the third round. Such a study is recommended. Our analysis suggests that some contestants may deserve to qualify for the top 100, for example contestants who answered all items attempted correct, subject to some lower cut-off.

From our analysis it can be recommended that rather than only taking the top 100 in the existing scoring approach (ZM), it may be advisable to consider all contestants who ended up in the top 150 or 200 in all four statistical applications. These contestants can be thought of as yielding consistent evidence for their selection.

**TABLE 12:** Comparison of three individuals, selected in top 100 under some criterion.

| Code | Description | Rank | Case | Total | Maximum Score | Items | Location | *SE* |
|---|---|---|---|---|---|---|---|---|
| F | Fay: Grade 8, school type unknown, female, Asian, language (not South African) | 208 | MM | 14 | 20 | 20 | 0.953 | 0.521 |
| | | 187 | MMT | 14 | 19 | 19 | 1.230 | 0.577 |
| | | 84 | ZM | 14 | 20 | 20 | 0.971 | 0.525 |
| | | 108 | ZMT | 14 | 19 | 19 | 1.266 | 0.585 |
| G | Grant: Grade 8, private school, male, White, English | 14 | MM | 11 | 11 | 11 | 2.569 | 1.309 |
| | | 26 | MMT | 11 | 11 | 11 | 2.588 | 1.333 |
| | | >354 | ZM | 11 | 20 | 20 | 0.232 | 0.493 |
| | | >354 | ZMT | 11 | 19 | 9 | 0.367 | 0.534 |
| H | Henry: Grade 8, government school, male, White, Afrikaans | 31 | MM | 12 | 13 | 13 | 2.156 | 0.950 |
| | | 48 | MMT | 12 | 13 | 13 | 2.113 | 0.954 |
| | | 261 | ZM | 12 | 20 | 20 | 0.469 | 0.499 |
| | | 190 | ZMT | 12 | 19 | 19 | 0.650 | 0.546 |

**TABLE 13:** Item comparison of two contestants with a common score 14 (70%).

| Item | Location | Adri | | Bets | |
|---|---|---|---|---|---|
| | | Response | Weighted / 100 | Response | Weighted / 100 |
| 16 | –1.529 | 1 | 6 | 1 | 6 |
| 6 | –1.265 | 1 | 5 | 1 | 5 |
| 4 | –1.183 | 1 | 4 | 0 | -1 |
| 13 | –0.844 | 1 | 5 | 1 | 5 |
| 1 | –0.665 | 1 | 4 | 1 | 4 |
| 3 | –0.596 | 1 | 4 | 1 | 4 |
| 19 | –0.518 | 1 | 6 | 1 | 6 |
| 14 | –0.500 | | 0 | 1 | 5 |
| 2 | –0.331 | 1 | 4 | 1 | 4 |
| 17 | –0.276 | 1 | 6 | 1 | 6 |
| 9 | 0.132 | 1 | 5 | 0 | -1 |
| 7 | 0.383 | 1 | 5 | 1 | 5 |
| 15 | 0.429 | | 0 | 1 | 5 |
| 8 | 0.510 | 1 | 5 | 1 | 5 |
| 11 | 0.555 | 1 | 5 | 0 | -1 |
| 5 | 0.923 | 1 | 4 | 1 | 4 |
| 18 | 0.986 | | 0 | 0 | -1 |
| 20 | 0.994 | | 0 | 1 | 6 |
| 12 | 1.390 | | 0 | 0 | -1 |
| 10 | 1.405 | | 0 | 0 | -1 |
| **Total score** | | **14** | **68** | **14** | **64** |

From the analysis it appears that the test is functioning adequately. The different approaches have slightly different effects on the difficulty levels of the items, but have a greater effect on the selection of contestants. The various analyses resulted in different subsets of contestants falling within the top 100. The reason for the differences rests on different views of underlying proficiency and therefore the differing approaches to missing data. A rationale may be made for the selection of one data set; however, a composite arrangement may offer advantages.

The empirical evidence for the ranking must be critically examined as other factors may contribute to ranking, for example non-coverage in some schools of related curriculum elements.


# Acknowledgements

We would like to express our gratitude to the South African Mathematics Foundation (SAMF) for providing access to the data, to Anna James for initial data cleaning and analysis and to Estelle Botha for the initial technical editing of the article.


## Competing interests

The authors declare that they have no financial or personal relationships that may have influenced the writing of the article.

## Authors' contributions

All four authors were responsible for the conceptualisation of the study, for writing sections of the initial draft and for contributions to the data analysis. C.L. conducted the Rasch analysis and took responsibility for the manuscript as a whole. J.E. was project leader and played an important role in describing the function of mathematics Olympiads, as well as the critique. He commented on all drafts. V.S. was involved in the writing of the missing data section, in addition to commenting on drafts of the article. T.D. played an important role overseeing the Rasch analysis and its interpretation. He commented on several drafts.

# References


Andrich, D. (1982). An index of person separation in Latent Trait Theory, the traditional KR-20 index, and the Guttman scale response pattern. *Educational Research and Perspectives*, *9*(1), 95–104.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: SAGE Publications.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms. In E.V. Smith Jr., & R.M. Smith (Eds.), *Introduction to Rasch Measurement* (pp. 143–166). Maple Grove, MN: JAM Press.

Andrich, D. (2006). *On the fractal dimension of social measurements I*. Perth: Pearson Psychometric Laboratory, University of Western Australia.

Andrich, D., Marais, I., & Humphry, S. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics*, *37*(3), 417–442.

Davis, G.A., Rimm, S.B., & Siegle, D.B. (2011). Education of the gifted and talented (6th edn.). Boston, MA: Allyn and Bacon.

Dunne, T., Long, C., Craig, T., & Venter E. (2012). Meeting the requirements of both classroom-based and systemic assessment of mathematics proficiency: The potential of Rasch measurement theory. *Pythagoras*, *33*(3), 1-16. http://dx.doi.org/10.4102/pythagoras.v33i3.19

Engelbrecht, J., & Mwambakana, J. (2016). Validity and diagnostic attributes of a mathematics Olympiad for junior high school contestants. *African Journal of Research in Mathematics, Science and Technology Education*, *20*(2), 175–188. http://dx.doi.org/10.1080/18117295.2016.1190211

Kenderov, P.S. (2006). Competitions and mathematics education. In M. Sanz-Solé, J. Soria, J.L. Varona, & J. Verdera (Eds.), *Proceedings of the International Congress of Mathematicians* (pp. 1583–1598). Madrid: IMU.

Ludlow, L.H., & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement,* 59(4), 615–603.

Mallinckrodt, C.H., Sanger, T.M., Dubé, S., DeBrota, D.J., Molenberghs, G., Carroll, R.J., et al. (2003). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry*, *53*(8), 754–760.

Masters, G.N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, *25*(1), 15–29.

McKnight, P.E., McKnight, K.M., Sidani, S., & Figueredo, A.J. (2007). *Missing data: A gentle introduction*. New York, NY: Guilford Press.

Parker, M., & Leinhardt, G. (1995). Percent: A privileged proportion. *Review of Educational Research*, *65*(4), 421–481.

Pigott, T.D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, *7*(4), 353–383.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded edn. with foreword and afterword by B.D. Wright. Chicago, IL: The University of Chicago Press. (Original work published 1960)

Ridge, H.L., & Renzulli, J. (1981). Teaching mathematics to the talented and gifted. In V. Glennon (Ed.), *The mathematics education of exceptional children and youth, an interdisciplinary approach* (pp. 191–266). Reston, VA: National Council of Teachers of Mathematics.

Sijtsma, K., & Van der Ark, L.A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, *38*(4), 505–528.

Tennant, A., & Conaghan, P.G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research*, *57*(8), 1358–1362.

Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of Operations and Management*, *24*(1), 53–62.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* New Jersey: Lawrence Erlbaum Associates.