Australian Journal of Adult Learning
Volume 46, Number 2, July 2006

An evaluation of a formal professional examination in adult continuing education

James A. Athanasou
Faculty of Education
University of Technology, Sydney

This study provides a framework for the evaluation of assessments that may be used in adult continuing education. It provides an example of the analysis of an examination for 33 solicitors seeking specialist accreditation. Resampling was used to generate a group of 1000 results, and responses were analysed using a Rasch model. Results indicated a select and capable group of candidates for whom many items in the assessment were redundant. A five-step general model for evaluating formal assessments in adult education is outlined.

Introduction

Quite appropriately, adult education is not viewed as a field in which formal educational assessments are dominant let alone as a

quantitative area of research (see English, 2005; Knowles, Holton & Swanson, 1998). Nevertheless, there are pockets of continuing adult education and training which involve formal, high stakes assessments and that are often overlooked as components of this diverse field. One such area in Australia is the specialist accreditation of solicitors that recognises their expertise in specific areas, such as: advocacy, business law, commercial litigation, criminal law, employment law, family law, immigration law, local government and planning law, mediation, personal injury, property law, taxation or wills and estates.

The specialist accreditation of solicitors is a nationwide endeavour but one undertaken independently at the state level by the various law societies. For instance, the Specialist Accreditation Scheme in New South Wales was established in 1992 and today there are around 1400 accredited specialists. In 2005 around 122 applicants sat for the different specialist accreditations. Accreditation involves a three-phase assessment, including preparation of a mock file for a complex matter, an exam and an oral assessment in the form of either a peer interview or simulated client interview (Gonczi, Hager & Palmer, 1994).

The purpose of this paper is to analyse the performance of one cohort on this professional assessment. Typically, however, only a small number of practitioners seek specialist accreditation and these are by definition already a distinct group. Thus the evaluation of this formal assessment is hindered at the outset by a limited and select data set. Notwithstanding this limitation, the paper will also demonstrate to adult educators how resampling can be used to overcome the problem of small sample size and it will also apply the method of Rasch scaling to assist in the evaluation of this formal assessment. This is the first application of these methods in this field and a detailed description will be provided for the reader. Some aspects of this report may appear quantitative but the reader is assured that a statistical background is not essential. Where possible, straightforward

descriptions will be used and the interested reader will be referred to other sources. The results will have implications wherever formal assessments are used in adult education and training contexts.

The Rasch measurement approach described in this paper deals with the development of any assessment. It is consistent with the philosophy of having a clear definition of the construct being assessed (see for example Wilson, 2006) and recognises that valid inferences from the results constitute the essential quality of a sound assessment. Rasch analysis is ideally suited for contexts where there is a need for qualitative accuracy in describing a person's performance on a set of tasks. The emphasis is not on the total score but moving towards results that are descriptive and meaningful, conceptually coherent and structured. Having constructed an instrument (that is, developed a method) that is appropriate for the context, the next step in a sound assessment process is – as Wiggins (1998) quite rightly noted – the appropriate educative use of an assessment (*cf*. Athanasou & Lamprianou, 2002).

The specific focus of this paper is to evaluate the Personal Injury Law exam of the Specialist Accreditation Board of the Law Society of New South Wales. It is one component of a large continuing education program and this particular adult professional examination happens to be the most popular field for accreditation. It covers the fields of workers' compensation for injury, motor vehicle accident compensation and general liability for injury.

Description of the examination

The Personal Injury Law examination is a three-hour written paper (20 minutes reading time) in two parts: Part A comprises two essay questions worth 20 marks each and Part B comprises 20 short questions worth 3 marks each. This is a closed book examination and the identity of the candidates is not revealed.

The examination is constructed by assessment panels who are expert in the particular area of law. These subject-matter panels also design the mock file and conduct the peer interviews. As in many other professional fields, there is no formal training in educational assessment but there is considerable practical expertise in the content area. In-service courses for assessors have emphasised an approach (Gonczi, Hager & Palmer, 1994) in which different sources of evidence contribute to a judgement about professional competence. The professional knowledge within a specialty has always been considered a key component of this model and is assessed by a formal examination. In this respect, it serves as a useful prototype for other formal assessments in adult continuing education.

A typical Part A question deals with hypothetical cases about which the candidate is asked to give preliminary advice. This involves a consultation by two new clients and the candidate is asked to give some brief preliminary advice about each case. For each case the candidate is required to explain (with reasons): the relevant parties to any possible action and the causes of action which should be considered; the issues which are most likely to be contentious; the defences, if any, which may be raised; the statutes, if any, which may be relevant; the court or tribunal which is most likely to be appropriate; and what steps or enquiries might be made immediately and before filing a claim on the client's behalf.

Part B of the examination consists of 20 questions selected from a pool of questions previously forwarded to the candidates. Answers are intended to be brief and should include references to relevant statutory provisions where applicable. Two sample questions are: "Is it possible for a court to make a 100% reduction of damages in respect of a plaintiff's contributory negligence?" and "In a claim under the *Compensation to Relatives Act 1897*, what is the effect of contributory negligence on the part of: (a) the deceased relative; and (b) the claimant?" A description of the results for all 33 candidates from the

2005 Personal Injury Law examination is provided in the following section.

Analysis of the results of candidates

Typically, the assessment panel obtains the results by adding together marks for the various components and sets a pass mark of 50%, as is common practice in most secondary and tertiary sectors in Australia. In 2005, the final scores ranged from 56 to 85 out of a possible 100, with a mean score of 71. By all accounts, this was deemed a competent group as everyone passed (see Figure 1 for a distribution of the overall results).

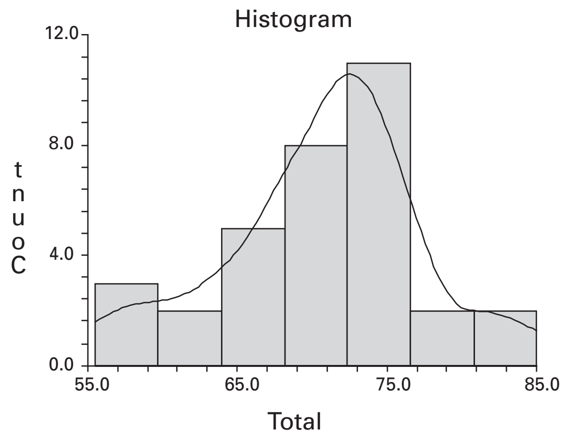

Figure 1: Distribution of scores on the Personal Injury Law examination

Rasch measurement

Analysis of raw scores, however, is not an adequate approach to educational assessment. Scores hide as much as they reveal. For example, the average score on an assessment reflects the competence of the group but the competence of the group is dependent on the difficulty of the assessment. Therefore, both difficulty and competence

are intertwined. Secondly, scores are not real units of learning or competence in the sense that seconds, metres, kilograms and litres have real world equivalents. Thirdly, they lack the fundamental aspect of additivity (Michell, 1994, 1997). For instance, it is certainly the case that scores can be added arithmetically but in reality they do not represent equal units. At best, scores only give us a vague sense of the extent of performance and are mainly useful for describing those with extremely high or low scores. We have known for many years that the units of ability represented by these numbers are not equal and this has practical implications. In this case it would require more ability to move from the extremes of 84 to 85 or 56 to 57 on this examination than it would take to move from 70–71, which is around the average. Finally and as a corollary, scores usually reveal little about the competence of the person in terms of the specific tasks that he/she is capable of undertaking correctly.

The Rasch methodology provides a way of overcoming these obstacles. It was developed by the Danish mathematician Georg Rasch (1960) principally in relation to reading attainment tests. Rasch methods are now widely used in large-scale educational assessments such as the *Program for International Student Achievement* (PISA) or the *Third International Mathematics and Science Study* (TIMMS).

Rasch used the method of conditional probability to overcome the problem of the interdependence of ability and difficulty in assessments. In doing this, he also provided us with a measurement unit (the logarithm of a probability) that could be added. Thus, he satisfied the fundamental criterion of assessment with the additivity of units (see Athanasou & Lamprianou, 2002; Bond & Fox, 2001).

Rasch also measured ability (that is, competence) and difficulty on the same scale and in the same units, meaning that for the first time both ability and difficulty could be compared. Using Rasch measurement it is now possible to determine in advance whether someone is likely to

have the competence to undertake a task. Finally, Rasch also provided mathematical models against which each item, task or question in an assessment could be compared, that is, we can determine the fit of the task to the model.

The Rasch model is essentially very simple and logical (Baker, 2001; Wright & Stone, 1979). It states that the probability of answering a question correctly is a function of the difficulty of the task and the competence of the person (see Appendix 1 for a mathematical expression of the model). When the competence of the person is greater than the difficulty of the task, then there is a higher probability (not absolute certainty because we are human after all) of answering it correctly. When the competence of the person is less than that of the task's difficulty, then there is an increased likelihood of failing on the task. While this verbal description is straightforward, the underlying mathematics is more complex.

Typically, Rasch measurement takes a very large sample that has undertaken an assessment and describes the performance of the group on each of the tasks in the assessment process. Note again that the emphasis is not on the total score but on the response to each item, question or task (Baker, 2001). This is why it is also called item response theory and this model of responding can be tested. It is also possible to see whether each person responded in a way that was consistent with the model. Of course, our small sample of 33 experienced solicitors provides a major constraint in applying Rasch methods, but it is really quite typical of adult learning where we usually have small cohorts.

One way around this is to use the technique of resampling to create a larger sample with smaller standard errors. This is achieved by continuously sampling with replacement from the original sample (Agho & Athanasou, 2005). In this case I took small samples from the group of 33 and continued until a total sample of 1000 was achieved. While this might seem like some sort of statistical conjuring, it actually produces a sample which is much more likely to be representative of the original population (see Effron, 1979). The technique is also called "bootstrapping" because one literally lifts oneself up by his/her bootstraps. The process normally ceases when the statistic of interest stabilises its value and the calculated errors of measurement are reduced to a desired level.

Resampling is now a commonplace technique for replicating the original population. Essentially any small group that we have in adult education is from a larger potential population. If we continuously take thousands of small samples from our group, then we can come closer to replicating the larger potential population (see Simon 1999 for further details). With this brief introduction to both Rasch and bootstrapping/resampling, it is now time to turn our attention to applying these to this formal continuing education assessment.

Sampling with replacement

The results for all tasks were first recoded on a common scale from 0 to 5 for each question in Part A and also in Part B. (The essay questions, Part A: questions 1 and 2, that were marked out of 20 have been recoded as: 0=9; 1=10; 2=11–13; 3=14–16; 4=17–18; 5=19–20. The Part B questions that were marked out of 3 have been recoded as: 0=0; 1=.5; 2=1; 3=1.5; 4=2; 5=2.5–3.0.) The maximum total score on the test was now 110. Then a resampling of 1000 with replacement was undertaken to increase the size of the sample in order to eliminate the problems of a small initial sample size and to reduce any errors of estimation.

Final scores now ranged from 56 to 85 (all figures rounded to nearest whole number) out of a possible total of 110 with a mean score of 83 (95% confidence level of 83 to 82; standard deviation = 5). Generally, one could have reasonable confidence in a score of 82 as the average for this group out of a maximum 110. Looking at the distribution of responses in Figure 2, it implies that the group was of a high

standard. As one might expect, the results using resampling were now more evenly distributed around the average score than with the original sample of 33 (compare the shapes of Figures 1 and 2).

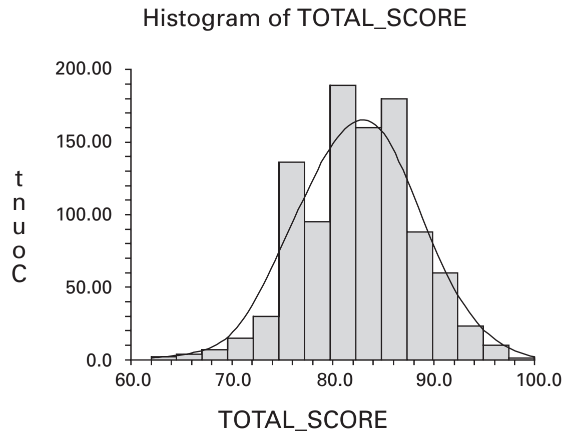

Figure 2: Plot of N=1000 for the total score

A Rasch item analysis

A Rasch analysis now takes the recoded scores and analyses them in terms of the contribution of the ability/competence of each person to their performance on each item or question. (The Rasch analysis was undertaken using both the QUEST (Adams & Khoo, 1994) and the RUMM (Andrich, Sheridan & Luo, 2004) programs.) To repeat the main point, the model that is being tested is that those with the highest competence should perform better on each question than those with lower ability. If this does not hold, then the item or task is not assessing what it is intended to determine.

For the reader who is concerned how the person's competence is determined, then in the absence of any other criterion, we are forced to use a person's total score on the exam as the initial estimate of ability. It is then possible to examine whether each item or question

fits this model using a range of measures derived from a Rasch analysis. Further details are explained below.

Item-ability map

The results of this assessment are summarised in the item-ability chart below (Figure 3). On the far left-hand side are the scores on this assessment which range from 4.0 (hard questions requiring high ability to answer them correctly) to -2.0 (easy questions requiring low ability to answer them correctly). The average on this Rasch scale is set at 0.0 but it can be transformed easily to any desired level (for example, 50 out of 100). In the language of Rasch measurement, these units are called logits (or log odds of probability). Typically they vary from +3.0 (items which are difficult or people with high competence) to -3.0 (items which are easy or people with low competence).

The series of Xs on the right-hand side is a distribution of the ability/competence of the candidates – it is like a chart that has been rotated 90 degrees. A general inspection reveals that competence on this assessment was fairly normally distributed with a slight tendency for a more average (0.0) than high average level of performance (>1.0). This is consistent with the selected nature of the group of candidates.

On the right-hand side and on the same scale as ability are located all the questions. These are expressed as a decimal. The first part of the decimal is the question number and the second part is the rating on that question. So 2.5 represents a rating of 5 on question 2. Closer inspection of this diagram shows that many items were well below the ability level of the group while some other questions were well above the ability level of the group. For instance, although hardly anyone failed the Part A (Question 1 and Question 2), it was difficult for them to obtain a rating of 4 or 5 on these questions. Questions below 0.0 failed to discriminate in terms of the dimension being assessed, namely, knowledge of personal injury law.

```
----------------------------------------------------------------------------
 4.0  High ability              |Difficult questions
                                |
                                |
                                |
                                |
                                |
                                |
 3.0                            |       2.5
                                |
                                |
                                |
                                |
                                |
                                |       1.5
                                |
                                |
 2.0                            |
                                |
                                |
                                |
                                |
                                |
                                |       1.4    2.4   19.5
                                |      15.5   18.5
                                |       5.5
 1.0                           X|
                             XXX|      17.5
                            XXXX|       3.5   13.5   16.5   20.5   21.5
                         XXXXXXX|       9.5   14.5
                  XXXXXXXXXXXXXX|       4.5    5.4    8.5   11.5   20.4
                   XXXXXXXXXXXXX|       7.5   12.5   15.4   18.4
            XXXXXXXXXXXXXXXXXXXX|       8.4    9.4   21.4
                     XXXXXXXXXXX|       3.4   16.4   19.4
                         XXXXXXX|       5.3    6.5   10.5   13.4   17.4
                               X|       4.4    8.3
 0.0                            |       6.4   14.4   18.3
                                |      13.3   17.3   20.2   21.3
                                |       3.3    7.4    8.2   10.4   11.4
                                |       1.3    4.3    5.1    5.2    6.1
                                |       2.3    4.1    4.2    9.3   19.3
                                |       3.1    3.2    7.1    7.2   16.3
                                |       9.1    9.2   17.1   17.2   18.2
                                |      13.1   20.1
                                |       8.1   12.4   18.1
-1.0                            |
                                |       1.2   16.2
                                |      14.3   15.3   19.2
                                |
                                |      22.2
                                |       2.2
                                |
                                |
                                |
-2.0 Low ability                | Easy questions
----------------------------------------------------------------------------
```

Each X represents 12 candidates

Figure 3: Item-ability map (N=1000)

On most tasks it was possible to score 3 or 4 out of 5 and still be well below the average level of competence of this select group (items 1–4, 6–7, 9–19, 21). Inspection of these items indicated that possibly a layperson with a nodding acquaintance of personal injury law, such as an insurance claims manager, let alone someone assumed to have a specialist knowledge of personal injury law, might even be able to respond with a partially correct answer to some questions (for example, 'What is meant by the term "vicissitudes of life"? How does the court take into account the vicissitudes of life when assessing damages for future economic loss?'). It is likely that these questions are redundant for this group. Of course, this may not be a problem if each item or question is intended to work towards a specific level of competence and to use this exam in the spirit of criterion-referenced assessment. By the same token, it is hardly necessary to ask questions that are well below the ability of the group as they do not add greatly to the level of knowledge about a person's competence. A numerical index of the reliability of this assessment called separability can also be determined and this was considered moderate at around 0.69 and it would be improved by better selection of items. The next section goes beyond the overall assessment and considers the nature of each item.

Fit of the assessment to the Rasch model

Most of the items fitted the constraints of the Rasch model. That is, as the ability of the group increased, so did the probability of answering the question correctly. This is calculated using a statistic, INFIT(see Figure 4) and the infit measures were well within the accepted criteria of 0.7 to 1.3 (Adams & Khoo, 1994, p. 26).

```
INFIT       0.63      0.71      0.83      1.00      1.20      1.40      1.60
--------------+---------+---------+---------+---------+---------+---------+----
      item 1                                *|
      item 2                                 *
      item 3                                 *
      item 4                                 |*
      item 5                                *|
      item 6                                 |*
      item 7                                 *
      item 8                                *|
      item 9                                *|
      item 10                                *
      item 11                                *
      item 12                                *
      item 13                                *
      item 14                                *
      item 15                                *
      item 16                                *
      item 17                               *|
      item 18                                |*
      item 19                                |*
      item 20                                |*
      item 21                                *
      item 22                                *
--------------+---------+---------+---------+---------+---------+---------+----
```

Infit statistics should lie within 0.7 to 1.3 to show that they conform to the Rasch model

Figure 4: Item-fit statistics (N=1000)

One of the benefits of item-response theory and the Rasch model is that it allows the evaluation of the response to each item through an item characteristic curve. This is an S-shaped curve that shows the expected score associated with a given level of ability or competence. One would expect an increased probability of answering a question correctly as the ability or competence of the person increases. The actual curve can be compared against the theoretically expected curve based on the equation in Appendix 1 to see to what extent it fits or departs from the model. Thus, the essence of Rasch measurement is that it proposes a model and really many other models could also be proposed and tested, but by and large the Rasch model is efficient in explaining most of the results from an assessment (Wright & Stone, 1979).

Figure 5 shows two of the item characteristic curves – one for an item that does not fit the model well (Item 1) and the other for an item that fits the model just marginally better (Item 9). Item 1 was the Part A case consultation concerning "…Belinda, the widow of George, who died following complications from elective knee surgery" and item 9 was the question, "Briefly explain which provisions of the Civil Liability Act 2002 have particular relevance to a claim by a surfer against a surf club which involves injury suffered by the surfer while swimming between the flags at the beach patrolled on the day by the club?". Only two out of the 22 items are dealt with here due to limitations of space.

In Figure 5, the black dots represent the average performance of 10 ordered sub-groups within the sample of 1000. One is looking for a monotonically increasing score as the competence of the group increases. The line models the performance on the item, question or task and will vary from item to item. The higher the chi-square value that is shown above the chart, then the better will be the fit of the group's performance to the Rasch model. Neither of these items is an especially good fit to the Rasch model although Item 9 is marginally better than Part A Question 1. (The full set of 22 item characteristic curves is available freely from the author upon request, together with all the datasets and the QUEST and Rumm outputs.)

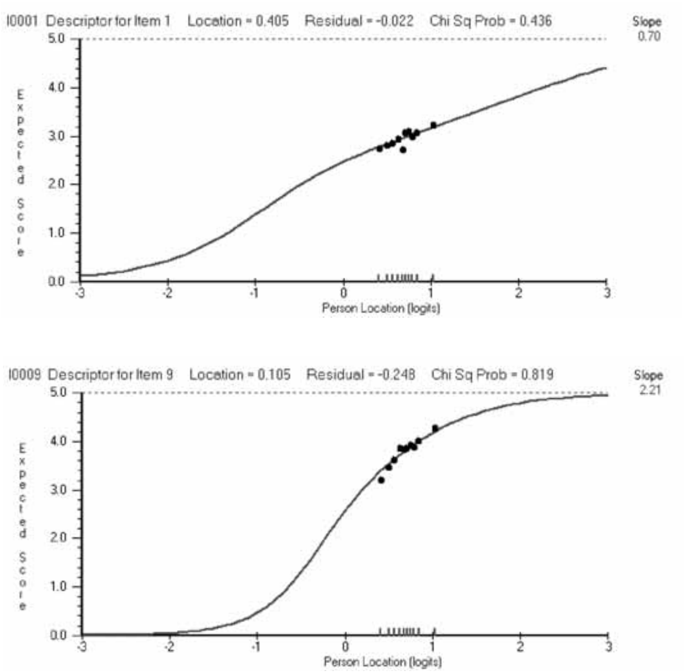

Figure 5: Item-characteristic curves for item 1 and item 9

The next chart that assists in the analysis of each item is the category probability curve. (The full set of category probability curves is also available from the author upon request.) The category probability curves may appear complex at first glance but they show five curves for the scores 0 to 5 for each item. The curve for each score shows the probability of answering correctly at each ability level. As expected, item 9 is clearly the better of the two items. If one follows the line for a score of 0, then there is a very high probability of scoring 0 with an ability of -3 logits, but by the time one reaches 1 logit, there is no probability of scoring 0. For a score of 5, there is almost no probability of scoring 5 up to -0.5 logits; it then increases rapidly until, with 3 logits of ability, it is almost certain one will score 5. Each of the scores can be traced and the respective probability of obtaining

that score read from the left-hand side and the ability level can be read from the bottom axis. By way of contrast, item 1 presents a less clear pattern.

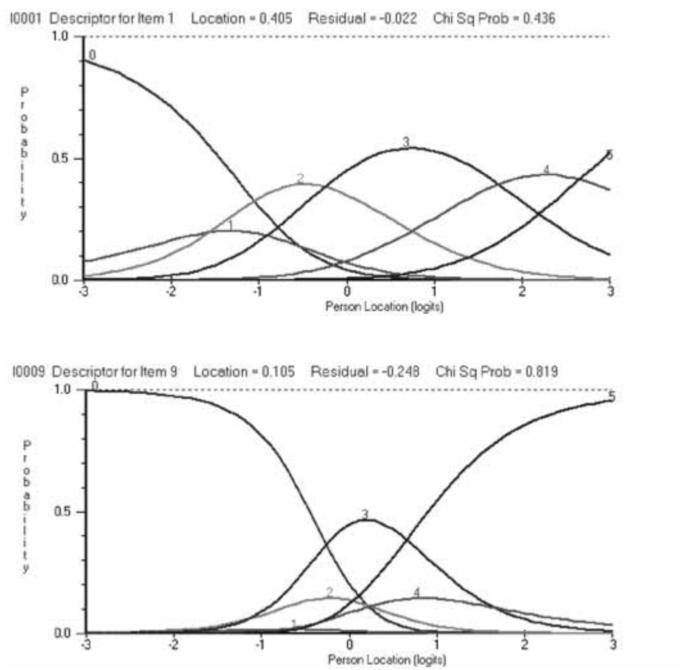

Figure 6: Category probability curves for item 1 and item 9

Concluding comments

The Rasch model is ideally suited to the analysis of competence on individual tasks rather than merely being a way of scoring assessments. In one sense, it provided an x-ray of the performance of this group of 33 adults on each task and it is much more consistent with a diagnostic adult education focus than previous ways of dealing with the results of assessments through scores or subjective judgements. These comments would also apply to questionnaires, surveys and attitude scales (see, for example, Athanasou, 2001).

The evidence that was obtained from this analysis now allows one to undertake an overhaul of the assessment in order to meet the needs of all stakeholders. For example, it was pointed out earlier that some items and tasks (items 1–4, 6–7, 9–19, 21) were not as useful as they were imagined to be at the outset. The content might have been relevant to the specialty but they did not provide information about competence. Consequently, one advantage of the application of Rasch measurement is in constructing or re-designing an assessment to ensure that every item and task in an assessment is a component of the construct or competence being assessed. Assessment panels have now been informed about the fact that the overall assessment was far too easy, that some items were redundant, and that other items were not providing helpful information about the level of knowledge of the candidate. Another advantage of the Rasch analysis is that the performance of each person can be examined or evaluated qualitatively. This has not been dealt with in this paper, but it involves examining the extent to which each person's pattern of results fits the Rasch model. It allows for a descriptive and interpretative view of what is happening with each individual (for example, tiredness, guessing, inconsistent patterns, gaps in their knowledge or problems with the structure of their competence).

There are some limitations of the Rasch model. The major restriction is that the assessment tasks for some reason or other may not fit the Rasch model which assumes that responses are a function of the specific characteristic of a person and the level of the task. The Rasch model is quite flexible and applicable across many circumstances, but it is conceivable that the observations may not fit and some other model should be developed. Another limitation is that some performances may be difficult to describe in terms of being right or wrong or along some scale. There are partial credit and rating scale Rasch analyses that can be used, but again it is conceivable that in some circumstances it will be difficult to describe or order performance. A further limitation is that the original conception

of the construct being assessed may not be coherent or structured. In some instances, it will be difficult to operationally define the construct, thereby making it impossible to assess. However, the overriding advantage of the Rasch approach is that it proposes a general model to account for responses and then sets out to test that model. That is both good science and good assessment practice.

In addition to providing an example of how a formal assessment might be analysed, one other objective of this paper has been to provide readers with a pragmatic framework for dealing with assessments in adult education, continuing education and training contexts. Without the resampling procedure, it would have taken some 30 years to amass a Personal Injury Law group equivalent in size to our bootstrapped sample of 1000. In the case of other specialties with only a handful of candidates, the evaluation of professional examinations has relied upon subjective judgements. The Rasch model provides a diagnostic and criterion-referenced focus rather than implicit and subjective norm-referenced approaches to analysing results within most educational settings (see Athanasou & Lamprianou, 2002).

Consistent with the approach outlined in this paper is a proposed framework for evaluating professional examinations in adult education. This is set out in Figure 7. It involves administering an assessment to a small group (N<250); coding the responses onto a common scale; resampling with replacement to produce a bootstrapped sample of size of 1000 or even better up to 10000; applying a Rasch analysis to each item to determine its strengths and weakness as well as its adherence to the model; and finally, applying a Rasch analysis to determine whether each individual's performance conforms with an expected pattern. The last stage has not been dealt with in this paper but it is available as a standard output from many of the Rasch analysis programs.

It is recognised that the application of resampling and the Rasch model may not have immediate application for most readers since they are not involved in formal assessments, but the potential application for surveys and questionnaires in adult education must not be overlooked. The same framework that was outlined in Figure 7 still applies. There may be a view that such approaches are suited to educational measurement contexts and hardly applicable to adult learning. In my opinion, this would represent a narrow view of measurement and a very limited perspective on the diverse field of adult education. Probably, the practices of educational assessment and educational measurement are much more conceptual and qualitative than appear at first glance (see Wilson, 2006).

Finally, Rasch approaches are consistent with assessment for learning. In recent years, there has been a distinct change in the ways in which educational assessment has been viewed. Previously, the summative aspects of assessment dominated educational discourses, but lately the emphasis has altered towards formative assessment as a component of curriculum and instruction. For instance, Wiggins (1998) described a view of assessment in which the aim "....is primarily to *educate* and *improve* student performance, not merely to *audit* it" (p. 7; italics are from the original quotation). Although the audit aspect has been emphasised in this paper, it is also recognised that certification *per se* does not exclude an approach to assessment that is both educative and informative. Without the necessary technical accuracy, however, neither the certification nor the educative objectives of an assessment will be achieved.

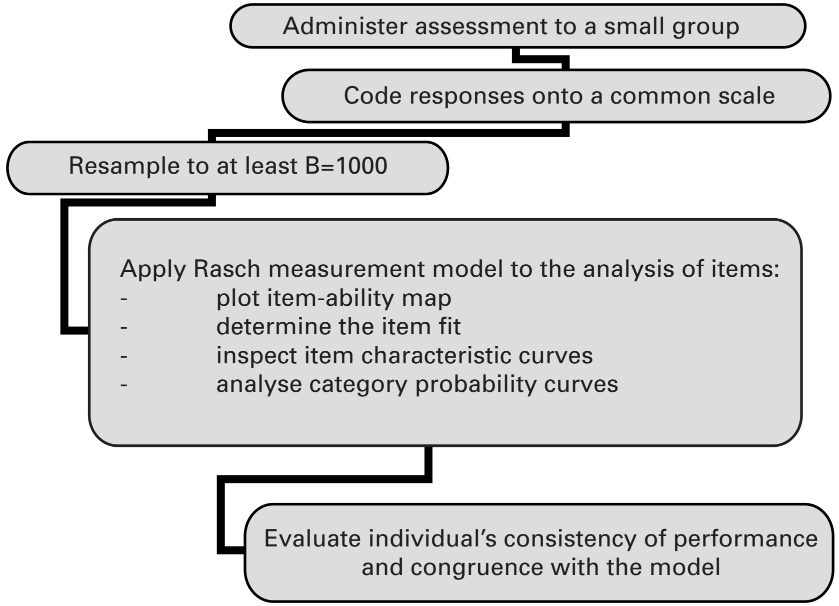

Figure 7: A framework for the evaluation of formal assessments

Acknowledgement

The cooperation of the Specialist Accreditation Board together with the assistance of Kingsley Agho, University of Sydney and the helpful comments of two anonymous reviewers are gratefully acknowledged.

References

Adams, R.J. & Khoo, S.T. (1994). *Quest: The interactive test analysis system*, Camberwell, Victoria: Australian Council for Educational Research.

Agho K.E. & Athanasou J.A. (2005). 'Software demonstration of the bootstrap method of Rasch conditional item difficulty estimation', *Rasch Measurement Transactions*, 19(2), 1022–3.

Andrich, D., Sheridan, B.E. & Luo, G. (2004). *Rasch unidimensional measurement models (RUMM): A windows based computer program,* Perth: Murdoch University.

Athanasou, J.A. (2001). ‘Analysis of responses to vocational interest items: A study of Australian high school students’, *Journal of Career Assessment*, *9*, 61–79.

Athanasou, J.A. & Lamprianou, I. (2002). *A teacher’s guide to assessment*, Sydney: Social Science Press.

Baker, F.B. (2001). *The basics of item response theory* (2nd ed.), Maryland: ERIC Clearinghouse on Assessment and Evaluation.

Bond, T. & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*, Mahwah, NJ: Lawrence Erlbaum.

Efron, B. (1979). ‘Bootstrap methods: Another look at the jackknife’, *Annals of Statistics*, *7*, 1–16.

English, L. (2005). *International encyclopaedia of adult education*, New York, NY: Palgrave Macmillan.

Gonczi, A., Hager, P. & Palmer, C. (1994). ‘Performance based assessment and the NSW Law Society Specialist Accreditation Program’, *Journal of Professional Legal Education*, 12(2), 135–147.

Knowles, M.S., Holton III, E.F. & Swanson, R.A. (1998). *The adult learner* (5th ed.), Woburn, MA: Butterworth-Heinemann.

Michell, J. (1994). ‘Numbers as quantitative relations and the traditional theory of measurement’, *British Journal for the Philosophy of Science, 45,* 389–406.

Michell, J. (1997). ‘Quantitative science and the definition of measurement in psychology’, *British Journal of Psychology, 88,* 355–383.

Rasch G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*, Chicago: University of Chicago Press. (Originally published by The Danish Institute for Educational Research, Copenhagen, 1960).

Simon, J. L. (1999). *Resampling: The new statistics (2nd ed.)*, Arlington, VA: Resampling Stats Inc.

Wiggins, G. (1998). *Educative assessment. Designing assessments to inform and improve student performance*, San Francisco: Jossey-Bass.

Wilson, M. (2006). *Constructing measures: An item response modeling approach*, Mahwah, NJ: Lawrence Erlbaum.

Wright, B.D. & Stone, M.H. (1979). *Best test Design. Rasch measurement*, Chicago: MESA Press.

APPENDIX 1: The Rasch one parameter logistic model

The one-parameter model (also called the Rasch model) assumes that item difficulty is the only characteristic affecting candidate performance. The item characteristic curve for the one parameter logistic model is given by:

$$P_i(£) = \frac{e^{(£. - b_i)}}{1 + e^{(£. - b_i)}}$$

$P_i(£)$ is the probability that a randomly chosen examinee with ability £ answers item *i* correctly (an s-shaped curve with values between 0 and 1 over the ability scale).

b_i is the item difficulty parameter. It is the point on the ability scale where the probability of a correct response is 0.5. This varies typically from –3 to +3 when values are transformed so that the average is 0.

e is the transcendental number (2.1718).

About the author

James Athanasou *is Associate Professor in the Faculty of Education at the University of Technology, Sydney. He lectures in the area of educational measurement. His research focuses on career development and vocational psychology. He is the editor of the* Australian Journal of Career Development.

Address for correspondence

Faculty of Education, University of Technology, Sydney, PO Box 123, Broadway 2007
Email: Jim.Athanasou@uts.edu.au