FOCUS ON METHODOLOGY

Refining assessment: Rasch analysis in health professional education and research

M. K. Farlie[1], C. E. Johnson[2], T. W. Wilkinson[3] & J. L. Keating[1]

Abstract

Educators want to assess learners using assessment processes that provide valid measures of learner ability. An ideal assessment tool would include items that are appropriate for assessing the target attributes. Ideal assessment results would accurately differentiate learners across the spectrum of ability, determine which learners satisfied the required standard and enable comparison between learner cohorts (e.g., across different years). Similar considerations are relevant to researchers who are designing or revising methods used to gather other kinds of assessment data, such as participant responses to surveys or clinical measurements of performance. Analysing assessment scores using Rasch analysis provides information about scores and the nature of each assessment item, and analysis output guides refinement of assessment. However, few health professional educators have published research that includes Rasch modelling methods. It may be that health professional educators find the language used to describe Rasch analysis to be somewhat impenetrable and that this has, to date, limited engagement in exploring applications for Rasch. In this paper, we lay out an overview of the potential benefits of Rasch analysis in health professional education and research.

Keywords: Rasch analysis; assessment; measurement; learner ability; scale development

[1] Department of Physiotherapy, School of Primary and Allied Health Care, Faculty of Medicine, Nursing and Health Science, Monash University, Melbourne, Victoria, Australia

[2] Monash Doctors Education, Monash Health and Faculty of Medicine, Nursing and Health Science, Monash University, Melbourne, Victoria, Australia

[3] Department of Medicine, University of Otago, Christchurch, New Zealand and Education Unit, University of Otago, Christchurch, New Zealand

Correspondence
Melanie K. Farlie
Department of Physiotherapy, School of Primary and Allied Health Care
Faculty of Medicine, Nursing and Health Science
Monash University
47–49 Moorooduc Highway
Frankston, Victoria 3199
Australia
Tel: +61 3 9904 4524
Email: melanie.farlie@monash.edu

Analysing and refining assessment: What Rasch analysis can tell you

Whether you are creating a written test, a clinical assessment scale, a questionnaire or a specialisation examination, it is important to know that your assessment provides you with valid measurements of the attribute of interest. Although we will focus on assessment in health professional education, the key messages in this report apply to any kind of measurement that yields a performance score.

Assessment is integral to health professional education, serving as a process that both drives and evaluates learning. Assessment methods in health professional education include various forms of written tests and performance tests, and observation of clinical practice (Downing & Yudkowsky, 2009). Recent shifts in assessment include a greater emphasis on assessment "for" learning rather than solely assessment "of" learning (Schuwirth & Van der Vleuten, 2011). Within this evolution, and in the broader field of research into assessment, the imperative for quality and continual improvement in assessment practices persists (Norcini et al., 2018). Research into best practice in assessment also continues to evolve, for example, in the field of standard setting to determine pass or fail cut points in assessment (Mubuuke et al., 2017). Despite this, assessment practices in health professional education and research are more commonly based on traditional practices (Bearman et al., 2017), and assessment "literacy" is a relatively new field of study (Friesen & Mousavi, 2020). Rasch analysis presents important opportunities to review and refine your test and advance assessment quality (Boone et al., 2014; de Jong-Gierveld & Kamphuls, 1985).

Rasch analysis tells you whether or not your item set conforms to the requirements of fundamental measurement and guides construction of good quality test questions. We will call test questions "test items". In addition, Rasch analysis guides the development of a test that provides trustworthy evidence of learner ability. This is important regardless of how test results are used. The accuracy of assessment results affects decisions regarding which learners meet required standards for a pass or who might be eligible for entry into programs with limited places. In this introduction to Rasch analysis, we illustrate key concepts using the example of an educator who assesses student achievement with a written examination, but the messages are relevant to the development of any measurement scale for education or research.

Why Rasch?

Imagine that you teach research methods to final-year students in a health professional education program. Each year you set the exam with a combination of old and new items (questions). The written exam consists of 50 multiple-choice items. The answer to each item is scored as either right or wrong, with each correct answer scoring two points. The pass mark is 65%, and passing the subject is a necessary requirement for course completion. The grades are also used in determining eligibility for entry to an honours program. We might call the underlying ability you want to assess "research literacy". Your goals as an educator are supporting learners to achieve a satisfactory standard of research literacy, graduating practitioners who can apply the principles of evidence-based care in clinical practice and making fair selections when approving honours candidature. Your course has clearly specified learning outcomes, and you have designed your written exam items to explicitly target assessment of these learning outcomes.

This year, the mean test score was 78% (SD 8%). It had a score range of 59% to 98%, and six students failed. This is similar to scores for previous cohorts.

You have heard that Rasch analysis could tell you a great deal more about your exam, your learners and your course, but the Rasch overview you read last year was somewhat impenetrable. It seems like a lot of unnecessary work to investigate your assessment procedures when the university examinations board is very satisfied with the mean, standard deviation and score ranges that you currently give them.

In this paper we hope to whet your appetite with information that enables you to consider how Rasch analysis might improve the quality of your assessment procedures. We also signpost some further options if you are keen to learn more.

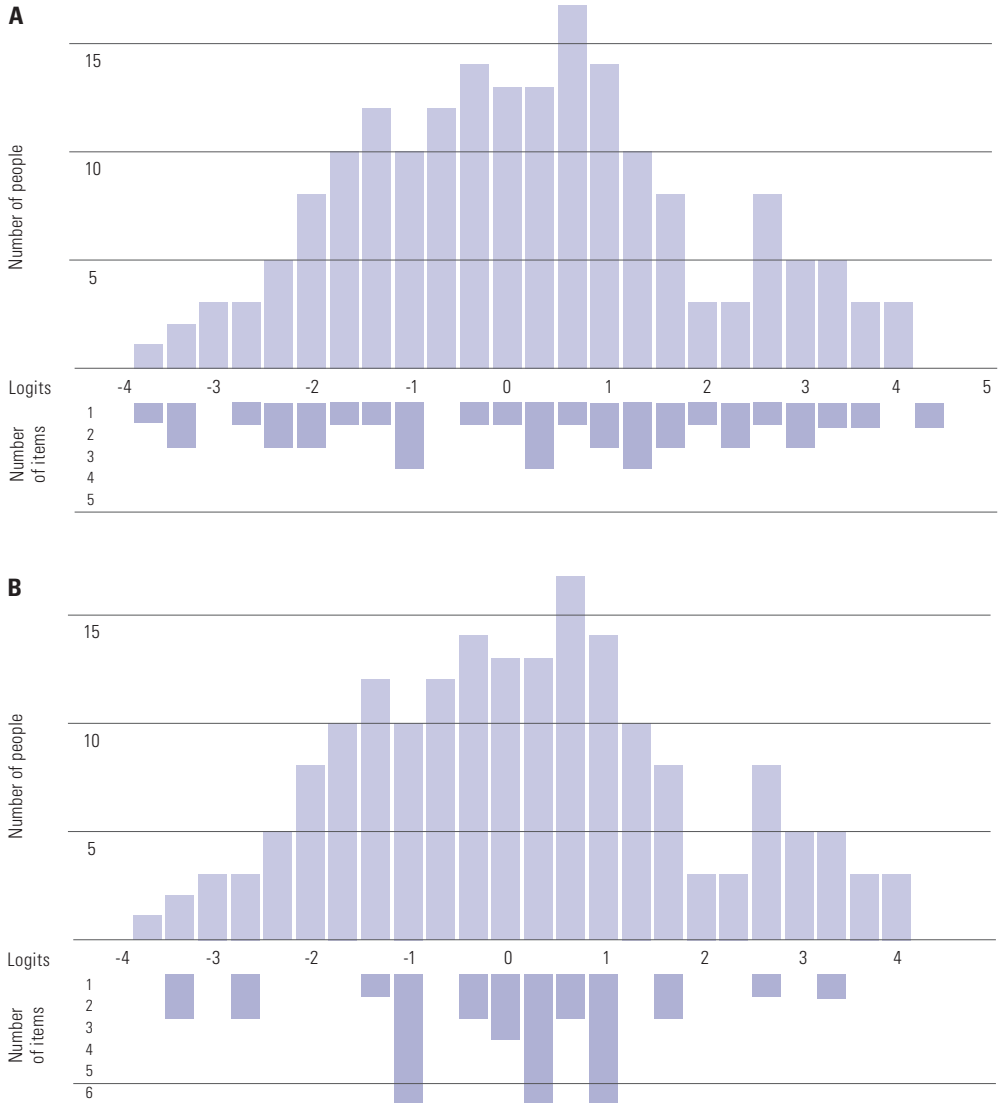
Some measurements are better than others

When you measure height with a ruler, you have confidence that the measurement is a consistent and accurate representation of height. You can trust that there are equal intervals between each centimetre mark on the ruler so that a difference of a centimetre between two people would be the same magnitude of difference, regardless of whether they were tall or short. When you test learners for research literacy, you are trying to measure them with a research literacy ruler. You might grade them with a score, but you do not know if the differences between each score on your ruler are equal (Andrich, 2011). Rasch analysis shows this to you and guides transformations of your measurement procedures to create measurements that behave like a ruler with standardised intervals between each score (Wright, 1977). When your test measurements are like this, meaningful comparison between learners and cohorts and valid interpretation of means and standard deviations are enabled (Merbitz et al., 1989; Wright & Linacre, 1989).

Figure 1

Person-Item Distribution Maps Showing "Person Ability"—"Item Difficulty" Distributions.

Note: Figure 1A shows a reasonable distribution of item difficulty with good coverage of learner ability, while Figure 1B shows poor coverage with several gaps in the items needed to estimate the ability level of learners at some difficulty levels and no items to classify the learners with the highest levels of ability.



How does Rasch analysis do these things?

Using total test scores, Rasch analysis estimates the ability level of each learner (person ability). Using the number of people who answered each item successfully, Rasch analysis estimates the difficulty level of each item on the test (item difficulty). It converts person

ability and item difficulty to the same units, called logits (pronounced lojit), and this enables person ability and item difficulty to be plotted on the same scale (Rasch, 1980).

Figure 1A shows such a plot, called a person–item distribution map. The "logit ruler" is shown running horizontally across the middle of the plot, ranging from logit -4 to +5. Above this logit ruler is the distribution of learner abilities ranging from around -4 to just above +4. The number of learners at each ability level is represented by the height of the columns. Below the logit ruler are the item difficulty estimates, i.e., the difficulty level of each item on your test. The depth of each column shows the number of items at each difficulty level. Items are assembled from left to right in order of ascending difficulty. Learners are assembled from left to right in order of ascending ability. The mathematics of Rasch create a logit ruler (Figure 1), where a learner with an ability level of, for example, 1 logit will have a 50% probability of correctly answering an item with a difficulty level of 1 logit. That learner will have a reducing probability of correctly answering items as the item difficulty rises (i.e., items further to the right, beyond a logit of 1) and an increasing probability of correctly answering items as the item difficulty decreases (i.e., items further to the left, below a logit of 1) (Rasch, 1980).

The person–item distribution map also displays information on person ability and item difficulty distribution. Figure 1A shows a test with a reasonable distribution of item difficulty, and this provides opportunities to test learners across the whole spectrum of ability. Only the most able learners are likely to answer the most difficult items correctly, and the less able learners have diminishing probabilities of answering all items that exceed their ability level.

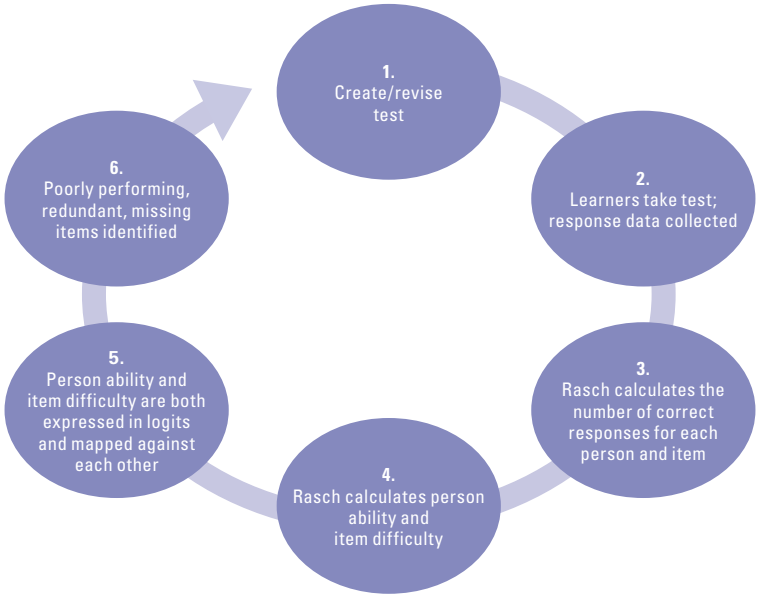
Now let's consider Figure 1B. Although we have the same distribution of learner ability, there are now some limitations in our ruler. There are a number of learners at the top of the scale, with an ability level that exceeds the highest level of item difficulty, and their true ability level cannot be assessed by the items in this test. Similarly, there are no items that are easy enough to estimate the true ability levels of some learners, i.e., they have lower ability than the items in this test can assess. The gaps between items, seen in the lower half of the plot in Figure 1B, indicate spaces in the scale with no ruler marks. Therefore, the ability levels of learners in these parts of the "research ability ruler" would be less precisely differentiated (for example, for learners with abilities between -1.7 and -2.5 logits). In addition, there are a lot of potentially redundant items with the same level of difficulty (for example, at -1 logit). We say potentially redundant, as some of these items may be key to assessment of learning requirements and, so, are not conceptually redundant. In addition, more items may be helpful in parts of the scale where greater precision is required. For example, you may want more items at key points in the scale such as at pass criteria, where more items may be usefully clustered to enhance precision. Nevertheless, Figure 1B represents a ruler with an irregular scale. Good targeting or coverage implies a comparable spread of person ability and item difficulty (Bond et al., 2021). Item difficulty is the unit of measurement on your research literacy ruler, and

Rasch output provides opportunities to improve the placement of "ruler marks" so that they are distributed evenly across the scale (Andrich, 2011).

Steps in the Rasch analysis process

Figure 2

Basics Steps in Using Rasch Analysis to Refine Tests



Rasch analysis involves a number of key steps, as illustrated in Figure 2 and described in more detail below.

1. Create an initial set of items. In our example, test items are created to assess learners across the "research literacy" spectrum. You create the initial set of test items by integrating various inputs, such as your knowledge and experience of the course material, other expert opinions and previous learner responses to items.
2. Administer the test. Learners take the test, answers are marked correct or not correct and these data are collected.
3. Analyse response data. The number of correct answers is calculated for each person and for each item.
4. Person ability and item difficulty are calculated.
5. Person ability and item difficulty are expressed in logits, which typically range from -4 to +4 (Wright, 1977). For those of you who want to know more, a series of explanatory memos derived from key scientific publications describing the mathematics underpinning these computations have been conveniently compiled at https://www.rasch.org/rasch.htm. A hierarchy for both item difficulty and person

ability is developed. Items and learners are arranged in order of increasing difficulty and ability. These calculations involve an iterative process until the likelihood of a person with known ability correctly answering an item with known difficulty can be most accurately predicted by the resultant model (Andrich, 1985; Kean et al., 2018). This model would predict that two learners with the same ability would have the same probability of correctly answering any given item. The higher the learner ability relative to item difficulty, the higher the probability of correctly answering the item (Rasch, 1980).

6. The resultant model is compared to the "ideal" Rasch model, and sources of variation are examined. The Rasch model predicts the probability of any learner correctly answering any item and compares that to the actual test data, i.e., it assesses "fit to the model" (Gustafsson, 1980). Analysis output identifies items that do not conform to the pattern predicted by the Rasch model (Rasch, 1980). It is here that the art of item design meets the science of Rasch modelling. When an item does not fit the model, the educator might ask questions such as "Is the item badly written?", "Is the material assessed by the item being taught in a confusing or ambiguous way?" "Might the item be assessing something other than research literacy?" Items can be merged, modified or removed, and the data can be re-evaluated with Rasch modelling to see if such changes indicate likely improvement in the fit to the model. The test data analysis can be rerun on subsequent learner cohorts, and further refinement of the test can occur at each iteration (Bond et al., 2021).

Hence, unlike statistics such as means and standard deviations that summarise and describe data, Rasch modelling also identifies opportunities to improve the test. The mathematics involved in comparing data properties to data that fit a Rasch model are relatively simple, but as there are a huge number of computations to complete, it is always done using software. Options in available software and ways to acquire software user skills are presented later.

Some fundamentals in Rasch analysis

Data that fit a Rasch model have particular properties, referred to as "specifications". These specifications are important to statisticians but also make intuitive sense to educators.

Monotonicity

Related to the idea of a hierarchy of item difficulty is the concept of monotonicity. To have confidence that the sum of test item scores accurately reflects the ability of your learners on the "research literacy" spectrum, the probability of a correct response to any given item should increase with increasing learner ability. This concept is called "monotonicity". When test items have dichotomous response options (e.g., yes/no; correct/incorrect), monotonicity is relatively simple to assess and refine. If there are more than

two response options (e.g., a 5-point scale from strongly disagree to strongly agree), there are additional complexities (for more information see "software and further study options").

Unidimensionality

When data "fit the Rasch model", the test scores are distributed in accordance with the model predictions. In other words, higher ability learners correctly answer more difficult items, and the probability that they will do this on any given item approximates the probability of success predicted by the model. Fit to the model is an indication of likely "unidimensionality" (the measurement of predominantly one construct), although additional confirmatory tests would be conducted. In our example, Rasch analysis assumes that "learner research literacy ability" and "item difficulty" are the only influences on the learner total scores. In other words, test scores are not influenced by other factors, for example, English language ability or knowledge of a specific medical condition.

When a test predominantly targets assessment of one specific ability, the total test score is a sufficient statistic for ability. Unidimensionality is supported when test scores are distributed in a way that the model predicts and further tested with a number of tests including post hoc tests run with Rasch software (beyond the scope of this paper). When unidimensionality is not supported by Rasch analysis, a possible explanation is that some of the items on the test are measuring something other than the ability you wish to measure. Developing quality test items that are unambiguous and that target specified learning outcomes is an art form in itself, and you will probably frequently find yourself revising items to improve them. Rasch software helps with this as it generates a number of test results that identify truant items. Reasons for items not fitting the model might include item construction (e.g., ambiguous wording) or that they are assessing "something else". It is then up to the test compilers to determine if that "something else" should remain in the test or not. Items that do not fit the model can be deleted, reconstructed, merged or reworded. When items are removed, Rasch analysis can be run again to test whether fit to the model is improved. When items are revised, you will have to wait until you run the test on the next cohort to assess the effect of the changes on model fit.

Some tests are designed to assess across a number of dimensions. This might occur in a single written exam that tests learners on anatomy, physiology, biochemistry and clinical decision making in the management of heart conditions. If most students had comparable ability across dimensions, it is not impossible that such a test might yield data that fit the Rasch model (here the underlying construct being assessed might be something like "cardiac conditions management ability"). It is also likely that learners might have strengths and weaknesses across test dimensions. For example, some students may be better at answering items in anatomy than in clinical decision making. In this example, the exam may actually be assessing two "dimensions", rather than one. Total

scores may not reveal unsatisfactory levels of ability in specific dimensions. Even in an integrated exam, Rasch analysis could be performed on the subset of response data for each dimension, and these sections could be subsequently refined.

Local item independence

"Local item independence" is the specification that correctly answering one test item should not be systematically associated with correctly answering another test item after the effect of learner ability and item difficulty have been statistically accounted for in the Rasch model. Local independence exists when the Rasch model of person ability and item difficulty explains all differences in the data. It would preclude the use of test items where a correct response to one item was a prerequisite to correctly answering a subsequent item. Tests run by Rasch software assess this by examining correlations between the part of each score that is not explained by the relationship between item difficulty and learner ability, i.e., "noise" versus "signal". After the effect of the underlying variable "research literacy" has been factored out, there should be no significant correlations between residual data that is not explained by the model. Where local dependency is identified, there are a number of strategies available for dealing with it, including deleting or modifying items, or combining dependent items. Importantly, there are valuable benefits to a test with local item independence. There are no impacts on exam utility if items are presented in random order, which is commonly required for e-assessment platforms. Items of comparable difficulty can be substituted from item banks to provide year-to-year test variation while maintaining a stable item hierarchy and difficulty level. In addition, with no test item dependency, individual exam items can be revised without affecting other items.

Group invariance

If your test effectively assesses research literacy ability, item difficulty and person ability, estimates will remain comparable regardless of the sample of learners or the sample of test items used in person–item calibration. This is termed "invariance". Invariance means that your scale unit is stable, no matter how you use it (one centimetre is a centimetre no matter what is being measured or which ruler you pick up) (Baghaei, 2010). With item invariance, even though item difficulty is estimated on a single sample, it will remain the same when used to test a sample subset or comparable sample. With person invariance, person ability estimates will remain the same when learners are tested using a subset of test items or comparable items. When there is invariance, score meaning and interpretations hold across populations and contexts (Bond et al., 2021; Messick, 1989). This enables the development of banks of items of known difficulty from which a valid test can be confidently assembled. It also enables accurate comparisons between learners in consecutive or different cohorts. It enables accurate equilibration of test standards across different settings, for example, pass marks for high-stakes assessments that represent the same level of ability across different years (Smith, 2001).

One way that invariance in learner ability is examined using Rasch is by differential item functioning tests (DIF). These are tests for systematic influences of specific learner characteristics on test scores (Bond et al., 2021). The person coordinating the program or unit is in the best position to consider factors that might influence test scores. Examples of DIF investigations for our "research literacy" sample might include English language literacy, learner age or gender. You would want confidence that the test items are not designed in such a way that create inequities in test difficulty. Lack of DIF (non-significant differences between subgroups) reinforces confidence in the invariance of the data (Hagquist et al., 2009).

Multifaceted Rasch analysis

The Rasch model, as presented in this article, describes a situation where learners answer questions that are marked correct or incorrect. In this test design, the score is influenced primarily by two factors: the learner's ability and the item difficulty. However, when there are more factors involved, a more advanced Rasch model is required. An example of this might be when a learner's performance is assessed by multiple examiners, such as during an OSCE or student selection interviews, where the severity of each examiner's rating may also influence the learner's score (Iramaneerat et al., 2008; Till et al., 2013). When there are additional factors such as these to consider, multi-faceted Rasch analysis can be used. However, the basic principles of analysis remain the same (Bond et al., 2021; Linacre, 1994).

Rasch applications in health professional education and research

In contrast to summary data, Rasch provides a method for refining the way learners are assessed (Wolfe & Smith, 2007a, 2007b). When data are analysed using Rasch modelling, the output provides you with statistics that support or challenge item monotonicity, unidimensionality, local item independence and invariance. Output does not validate or invalidate the usefulness of your test items but signposts items that may need modifying to improve the validity of the test (Smith, 2001; Wolfe & Smith, 2007b). Although Rasch converts learner ability to a logit score, logits can be converted back into percentage scores that better indicate ability than a simple summing of items answered correctly. The attributes of Rasch analysis have seen Rasch modelling used in the assessment of exams set within medical schools (Homer et al., 2012; Royal & Hedgpeth, 2017) and the development of medical specialisation exams (Scheuneman & Subhiyah, 1998; Yang et al., 2011). In addition to health professional education, there are broader applications of Rasch analysis in education, such as its use by international organisations like PISA, who designed the OECD's Programme for International Student Assessment (OECD, 2009).

Rasch skills have many other important research applications, as they can be applied to a broad range of assessment domains, for example, the development of performance tests such as the Balance Intensity Scale, a clinical test of balance exercise performance (Farlie

et al., 2019) or the Assessment of Procedural Skills in Physiotherapy, an assessment of procedural skill performance (Sattelmayer et al., 2020). Another example is the evaluation of competency to practise measured in the authentic workplace, such as the Assessment of Physiotherapy Practice (Dalton et al., 2011), which is now used in physiotherapy courses across Australia and New Zealand. Rasch has also been used in the development of methods to improve health educator practices, such as the Feedback Quality Instrument, which guides the observable behaviours of educators engaged in face-to-face feedback with learners in clinical practice (Johnson et al., in press). Rasch analysis enables critical evaluation of assessment based on traditional methods (Bearman et al., 2017) and a rich perspective on the interaction between learners and test items (Bond et al., 2021; Boone et al., 2014). It elevates the field of research into assessment and provides quantitative foundations for test modifications and improvements.

A body of educators and researchers who gain proficiency in using Rasch to routinely check on the likely validity of student exams would be of great value to universities. Likewise, these skilled scholars would make valuable contributions to the health professional education research community through increasing the transparency of assessment design processes.

Software and further study options

Software options for conducting Rasch analysis include RUMM2030 and Winsteps (both can be purchased). It can also be performed using R code, which is freely available online software. An excellent article by Robinson et al. (2019) discusses the pros and cons of RUMM2030 compared to R. These options all come with many resources that guide users in conducting tests and interpreting output (see, e.g., https://www.winsteps.com/courses.htm for WINSTEPS resources). The very useful text by Bond et al. (2021) provides detailed guidance on Rasch and its applications and also access to free online Rasch analysis software (ARMsteps and ARMfacets) with preloaded data and guidance on analysis. In addition, there are many online and face-to-face courses that can support a first-time Rasch analyst trying to navigate new terrain. Accessing the skills of an experienced Rasch analyst, either in one-to-one guidance or by taking a course, is a comfortable way to start your Rasch skill development. For those confident in basic research methods, purchasing or accessing software, entering your data and reviewing your results before consulting Rasch experts would provide an excellent combination of learning opportunities. However you go about it, there is a learning curve associated with gaining Rasch analysis skills. A list of useful resources that other educators have found helpful are in the recommended further reading list provided below.

Back to the case …

Let's return to your research literacy test results summary. In comparison to the mean, standard deviation and range values that you planned to submit to the board of examiners, your Rasch analysis provided you with a very different platform from which to consider the fairness and validity of your assessment procedures. For the subsequent year's exams, you have added items to catch the full spectrum of learner ability, removed some items that showed a DIF for English as a second language, combined some items that showed local dependency and identified some items that very able learners answered incorrectly (suggesting item ambiguity) and modified the item wording. You also identified some items you thought were easy, but that very few people answered correctly, and following this up, identified that the related content and activities were not provided to this cohort. Those items were subsequently removed from calculations of percentage scores. You plan to revise the curriculum with content that addresses this oversight for next year's cohort. You reviewed the items that fell below the ability level that converted to a score of 65%. You asked yourself if failure to correctly answer items above this level met the expected standards for practice, and you raised the pass mark to 74% to include other items that you would consider essential knowledge. Lastly, you noted that one area of the curriculum was difficult for all learners regardless of their levels of ability, and you resolved to revise the methods used in teaching and learning to see if this leads to improved scores in subsequent cohorts. You feel confident that the ability spectrum of next year's cohort will be assessed more accurately and better serve the high-stakes function of the examination. In addition, the incremental adjustments to your teaching and learning program should, over time, lead to a more research literate graduate. In parallel, after 3 years of Rasch analyses, with learner consent, you published the assessment modifications and rationale, highlighting how flaws in earlier versions of assessment had evolved under the Rasch spotlight. This inspired others who were teaching research literacy to analyse data from their own assessments, enabling system-wide, evidence-based review and revision of assessment practices.

Summary and conclusion

Rasch modelling offers health professional educators and education researchers many sophisticated insights into test properties that are not evident in means, standard deviations, ranges and percentiles. Analysis of fit to Rasch model specifications, and item and scale modifications that improve fit to the model, are the vehicles through which test procedures can be validated and refined. Rasch analysis can give you confidence in the validity of your assessment data and a pathway to systematically improving test design, curriculum content and learner ability. We propose that there is considerable opportunity

to improve assessment practices for educators and researchers who 1) understand the benefits of Rasch analysis 2) familiarise themselves with the key concepts described in the analysis output 3) apply Rasch analysis to data gathered using a test they have constructed and administered 4) refine test composition in response to the analysis output and 5) sequentially review test performance in subsequent learner cohorts. In parallel, a strong, Rasch-informed research platform could drive systematic advances in the development of test items across a broad spectrum of assessments or questionnaires relevant to the healthcare professional.

Recommended further reading

https://www.rasch.org/ provides an extremely comprehensive set of resources for exploring answers regarding Rasch analysis.

Bond, T., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Funding and conflicts of interest

The authors declare no conflict of interest and have no funding to declare.

Ethical approval

Not required

References

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. *Sociological Methodology*, *15*, 33–80. https://doi.org/10.2307/270846

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research*, *11*(5), 571–585. https://doi.org/10.1586/erp.11.59

Baghaei, P. (2010). An investigation of the invariance of Rasch item and person measures in a c-test (pp. 100–112). In R. Grotjahn (Ed.), *Der c-test: Beiträge aus der aktuellen forschung/the c-test: Contributions from current research*. Peter Lang.

Bearman, M., Dawson, P., Bennett, S., Hall, M., Molloy, E., Boud, D., & Joughin, G. (2017). How university teachers design assessments: A cross-disciplinary study. *Higher Education*, *74*(1), 49–64. https://doi.org/10.1007/s10734-016-0027-7

Bond, T., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge. https://doi.org/10.4324/9780429030499

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Dalton, M., Davidson, M., & Keating, J. (2011). The Assessment of Physiotherapy Practice (APP) is a valid measure of professional competence of physiotherapy students: A cross-sectional study with Rasch analysis. *Journal of Physiotherapy*, *57*(4), 239–246. https://doi.org/10.1016/S1836-9553(11)70054-6

de Jong-Gierveld, J., & Kamphuls, F. (1985). The development of a Rasch-type loneliness scale. *Applied Psychological Measurement*, *9*(3), 289–299. https://doi.org/10.1177/014662168500900307

Downing, S. M., & Yudkowsky, R. (2009). *Assessment in health professions education*. Taylor & Francis Group. https://doi.org/10.4324/9780203880135

Farlie, M. K., Keating, J. L., Molloy, E., Bowles, K.-A., Neave, B., Yamin, J., Weightman, J., Saber, K., & Haines, T. P. (2019). The balance intensity scales for therapists and exercisers measure balance exercise intensity in older adults: Initial validation using Rasch analysis. *Physical Therapy*, *99*(10), 1394–1404. https://doi.org/10.1093/ptj/pzz092

Friesen, D., & Mousavi, A. (2020, May 30–June 4). *Assessment literacy in higher education* [Paper presentation]. 2020 Conference of the Canadian Society for the Study of Education, London, Ontario.

Gustafsson, J. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *33*(2), 205–233. https://doi.org/10.1111/j.2044-8317.1980.tb00609.x

Hagquist, C., Bruce, M., & Gustavsson, J. P. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies*, *46*(3), 380–393. https://doi.org/10.1016/j.ijnurstu.2008.10.007

Homer, M., Darling, J., & Pell, G. (2012). Psychometric characteristics of integrated multi-specialty examinations: Ebel ratings and unidimensionality. *Assessment & Evaluation in Higher Education*, *37*(7), 787–804. https://doi.org/10.1080/02602938.2011.573843

Iramaneerat, C., Yudkowsky, R., Myford, C. M., & Downing, S. M. (2008). Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances in Health Sciences Education*, *13*(4), Article 479. https://doi.org/10.1007/s10459-007-9060-8

Johnson, C., Keating, J. L., Leech, M., Congdon, P., Kent, F., Farlie, M. K., & Molloy, E. (in press). Development of the feedback quality instrument: A guide for health professional educators in fostering learner-centred discussions. *BMC Medical Education*.

Kean, J., Brodke, D. S., Biber, J., & Gross, P. (2018). An introduction to item response theory and Rasch analysis of the Eating Assessment Tool (EAT-10). *Brain Impairment: A Multidisciplinary Journal of the Australian Society for the Study of Brain Impairment*, *19*(Special Issue 1), 91–102. https://doi.org/10.1017/BrImp.2017.31

Linacre, J. M. (1994). *Many-faceted Rasch measurement*. MESA Press. https://www.winsteps.com/a/Linacre-MFRM-book.pdf

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation*, *70*(4), 308–312. https://www.researchgate.net/profile/Charles-Merbitz/publication/20619750_Ordinal_scale_and_foundations_of_misinference/links/0fcfd50059d8897a5c000000/Ordinal-scale-and-foundations-of-misinference.pdf

Messick, S. (1989). *Validity* (pp. 13–103). In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education* (3rd ed.). Macmillan and American Council on Education.

Mubuuke, A. G., Mwesigwa, C., & Kiguli, S. (2017). Implementing the Angoff method of standard setting using postgraduate students: Practical and affordable in resource-limited settings. *African Journal of Health Professions Education*, *9*(4), 171–175. https://doi.org/10.7196/AJHPE.2017.v9i4.631

Norcini, J., Anderson, M. B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., Hays, R., Palacios Mackay, M. F., Roberts, T., & Swanson, D. (2018). 2018 Consensus framework for good assessment. *Medical Teacher*, *40*(11), 1102–1109. https://doi.org/10.1080/0142159X.2018.1500016

OECD. (2009). The Rasch model. In *PISA data analysis manual: SAS*. https://doi.org/10.1787/9789264056251-6-en

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (2nd ed.). The University of Chicago Press. https://doi.org/10.1177/014662168100500413

Robinson, M., Johnson, A. M., Walton, D. M., & MacDermid, J. C. (2019). A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/eRm/TAM/lordif). *BMC Medical Research Methodology*, *19*, Article 36. https://doi.org/10.1186/s12874-019-0680-5

Royal, K. D., & Hedgpeth, M.-W. (2017). The prevalence of item construction flaws in medical school examinations and innovative recommendations for improvement. *EMJ Innovations*, *1*(1), 61–66.

Sattelmayer, K. M., Jagadamma, K. C., Sattelmayer, F., Hilfiker, R., & Baer, G. (2020). The assessment of procedural skills in physiotherapy education: A measurement study using the Rasch model. *Archives of Physiotherapy*, *10*, Article 9. https://doi.org/10.1186/s40945-020-00080-0

Scheuneman, J. D., & Subhiyah, R. G. (1998). Evidence for the validity of a Rasch model technique for identifying differential item functioning. *Journal of Outcome Measurement*, *2*(1), 33–42.

Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher*, *33*(6), 478–485. https://doi.org/10.3109/0142159X.2011.565828

Smith, E. V., Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, *2*(3), 281–311.

Till, H., Myford, C., & Dowell, J. (2013). Improving student selection using multiple mini-interviews with multifaceted Rasch modeling. *Academic Medicine*, *88*(2), 216–223. https://doi.org/10.1097/ACM.0b013e31827c0c5d

Wolfe, E. W., & Smith, E. V., Jr. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I—Instrument development tools. *Journal of Applied Measurement*, *8*(1), 97–123.

Wolfe, E. W., & Smith, E. V., Jr. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. *Journal of Applied Measurement*, *8*(2), 204–234.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116. https://doi.org/10.1111/j.1745-3984.1977.tb00031.x

Wright, B. D., & Linacre, J. (1989). Observations are always ordinal: Measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*(12), 857–860.

Yang, S.-C., Tsou, M.-Y., Chen, E.-T., Chan, K.-H., & Chang, K.-Y. (2011). Statistical item analysis of the examination in anesthesiology for medical students using the Rasch model. *Journal of the Chinese Medical Association*, *74*(3), 125–129. https://doi.org/10.1016/j.jcma.2011.01.027