Online Readings in Research Methods
(ORIM)

Chapter 6

Is this a useful instrument? An introduction to Rasch measurement models

Celeste Combrinck

University of Pretoria

Recommended Citation

Combrinck, C. (2020). Is this a useful instrument? An introduction to Rasch measurement models . In S. Kramer, S. Laher, A. Fynn, & H. H. Janse van Vuuren (Eds.), Online Readings in Research Methods. Psychological Society of South Africa: Johannesburg. https://doi.org/10.17605/OSF.IO/ BNPFS

Is this a useful instrument? An introduction to Rasch measurement models

Celeste Combrinck

University of Pretoria

INTRODUCTION

Measurement is ubiquitous in modern-day life. Tests, questionnaires, surveys and various forms of assessment are used to measure individuals from early childhood and continue to be part of every phase of life thereafter. However, it cannot be assumed that instruments are valid and reliable for all contexts (Boone & Noltemeyer, 2017). The word *measurement*, as used by social and natural scientists, implies that precise criteria must be met before an accurate and useful measurement is achieved, which is not always taken into consideration when designing social science instruments (Fried & Flake, 2018). This chapter aims to introduce the concept and application of Rasch models via interpretation guidelines and an example. Rasch Measurement Theory (RMT) is a host of statistical models, nested within Rasch's (1960) initial development for dichotomous items, and is used to assess the internal functioning of items and instruments (Bond & Fox, 2015). The models produce statistics, which provide evidence of the reliability and validity of inferences derived from items and total scores (Boone, 2016). The statistics offer indications to refine and improve instruments. The applicability of an instrument to various populations can also be examined using Rasch statistics. In a multi-cultural assessment milieu, tests and questionnaires must be fair and valid for different groups of respondents (Pearce, 2018). All of this and more are indicated by Rasch statistics which assess the degree of accurate measurement. Social scientists need to measure a variety of cognitive, psycho-social and behavioural characteristics, and Rasch models provide the gauge of how well instruments function for the constructs being measured.

The South African perspective

Apartheid and colonialism left South Africa and its neighbours with a complex and somewhat disturbing history of assessment (Laher & Cockcroft, 2014). The use of assessments in South Africa to justify the misrepresentation, mistreatment and misclassification of groups led to an understandable backlash against testing and a weakening of psychometric development in the country (Claasen, 1997; Combrinck, 2018). Unfortunately, this means that the scientific design of instruments in the South African context has not grown as much as may be desirable.

Combrinck, C. (2020 Is this a useful instrument? An introduction to Rasch measurement models. In S. Kramer, S. Laher, A. Fynn, & H. H. Janse van Vuuren (Eds.), Online Readings in Research Methods. Psychological Society of South Africa: Johannesburg. https://doi.org/10.17605/OSF.IO/ BNPFS

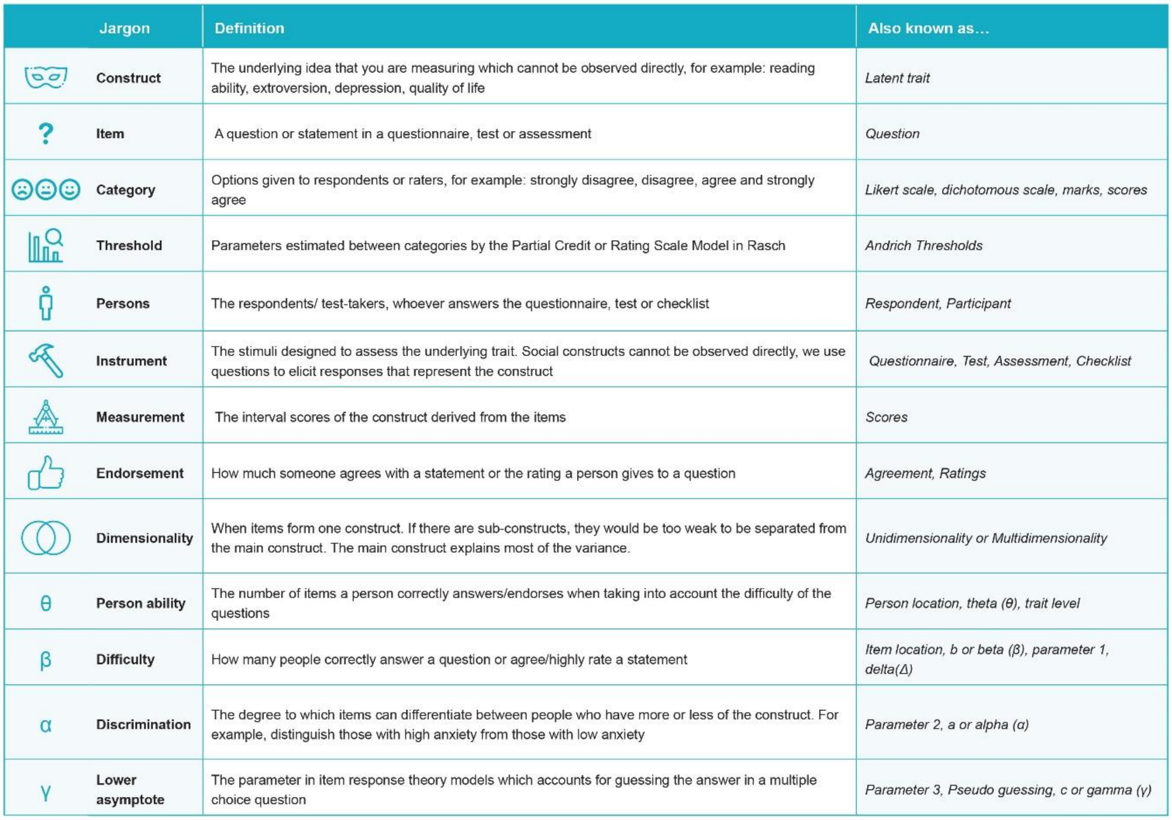
Developing countries have an even greater need for new instruments as we face challenges specific to our environment and the call to decolonise assessments (Barnes et al., 2018; Stead, 2002). What is measured, as well as how it is measured, is a crucial first step. The decolonisation of social sciences requires researchers from our context who are able to apply indigenous knowledge to psychometrics and reframe constructs for the diverse African milieu. To strengthen psychometrics, we need to recognise that:

1. psychological measurement is a science;
2. students and practitioners should be taught measurement science and application;
3. indigenous knowledge should be used to reconceptualise western constructs;
4. new constructs and theories need to be developed for an African context;
5. instruments from other settings should be used with great care; and
6. languages, cultures and the meaning of constructs for different groups should be investigated and addressed.

The decolonisation of assessment starts when we equip our early career researchers with the skills to evaluate instruments critically. Training in quantitative methodology and analysis is crucial. Researchers should also be able to develop their own assessments using the most robust psychometric techniques. Developing contexts may lack the intellectual capital to fully utilise modern psychometric models, which is why offering opportunities and resources to researchers is crucial (Combrinck, 2018; Laher & Cockcroft, 2013). This chapter is one of many open-access resources being made available to promote the measurement revolution in our context.

Jarring Jargon

Psychometric theory has many technical concepts. **Figure 1** contains some of the more frequently used terminology as well as explanations and alternative nomenclatures.



	Jargon	Definition	Also known as…
	Construct	The underlying idea that you are measuring which cannot be observed directly, for example: reading ability, extroversion, depression, quality of life	*Latent trait*
?	Item	A question or statement in a questionnaire, test or assessment	*Question*
☹😐☺	Category	Options given to respondents or raters, for example: strongly disagree, disagree, agree and strongly agree	*Likert scale, dichotomous scale, marks, scores*
	Threshold	Parameters estimated between categories by the Partial Credit or Rating Scale Model in Rasch	*Andrich Thresholds*
	Persons	The respondents/ test-takers, whoever answers the questionnaire, test or checklist	*Respondent, Participant*
	Instrument	The stimuli designed to assess the underlying trait. Social constructs cannot be observed directly, we use questions to elicit responses that represent the construct	*Questionnaire, Test, Assessment, Checklist*
	Measurement	The interval scores of the construct derived from the items	*Scores*
	Endorsement	How much someone agrees with a statement or the rating a person gives to a question	*Agreement, Ratings*
	Dimensionality	When items form one construct. If there are sub-constructs, they would be too weak to be separated from the main construct. The main construct explains most of the variance.	*Unidimensionality or Multidimensionality*
θ	Person ability	The number of items a person correctly answers/endorses when taking into account the difficulty of the questions	*Person location, theta (θ), trait level*
β	Difficulty	How many people correctly answer a question or agree/highly rate a statement	*Item location, b or beta (β), parameter 1, delta(Δ)*
α	Discrimination	The degree to which items can differentiate between people who have more or less of the construct. For example, distinguish those with high anxiety from those with low anxiety	*Parameter 2, a or alpha (α)*
γ	Lower asymptote	The parameter in item response theory models which accounts for guessing the answer in a multiple choice question	*Parameter 3, Pseudo guessing, c or gamma (γ)*

Figure 1. Rasch and IRT theory's technical terms and definitions (Source: Author)

Some of the terminologies are intuitive, for example, the term instrument. This is a collective name for questionnaires, tests, assessments and checklists. Terminology such as dimensionality, discrimination and the lower asymptote is more complex. The chapter offers examples to make the terms easier for the reader to interpret.

This chapter provides a guide on applying and interpreting Rasch theory by using two of the most popular software packages, Winsteps and RUMM. Practical knowledge of statistics is a prerequisite, including essential data processing skills. Other skills required include:

- understanding the concept of variables;
- the ability to capture/interpret data in a standard format of columns and rows;
- the ability to clean/screen data;
- knowing how to reverse code items;

- dealing with missing data[1];
- understanding basic statistical concepts such as significance and effect size, probability and hypothesis testing; and
- generally being able to manipulate data for further use and analysis.

These skills are beyond the scope of the current chapter. Introductory courses on statistics[2] can be used to enhance skills or online resources (see for example *Crash Course Statistics*[3]). Basic knowledge of statistics is a necessity for all consumers of social science, especially students. Psychologists and social science professionals need more than a basic understanding and should be well versed in quantitative methods. Familiarity with the levels of measurement, designated by Stevens (1946) as nominal, ordinal, interval and ratio is crucial. Stevens' (1946) work has been critiqued (see Burton-Jones & Lee, 2017; Michell, 2008; Wright, 1997) with regards to his definition of measurement: *measurement is the assignment of numerals to events or objects according to rules* (Stevens, 1959, p. 25). The problem here is that raw scores do not have equal interval measures, and assigning numbers to opinions, ratings and even test answers do not produce equal units or constant measurement between times and contexts. Assigning numbers to Likert scale ratings, test answers and other questionnaire responses implies that psychological constructs are quantitative (Michell, 2008). We need to acknowledge a paradox: when considering psycho-social constructs, we know that they are inherently qualitative. But for convenience of analysis, we sometimes ignore their qualitative nature and treat them as though they are fully quantitative. The erroneous treatment of constructs as though they are interval and numerical representations of complex human experiences leads to a host of problems, such as shallow, unreliable measurements and false interpretations. This does not change the fact that the defined levels of measurement remain used and accepted as doctrine. The responsibility lies with us; we must critically examine our operationalisation of the constructs we measure and

[1] The Rasch model can handle large quantities of missing data, both missing at random and planned missing data

[2] Statistics.com: https://www.statistics.com/

[3] https://www.youtube.com/playlist?list=PL8dPuuaLjXtNM_Y-bUAhblSAdWRnmBUcr

investigate the validity of our conclusions using both qualitative and quantitative methods. The paradox of measurement is also why more social scientists are recommending mixed methods to understand human constructs holistically, and more accurately (David, Hitchcock, Ragan, Brooks & Starkey, 2018; Pool et al., 2010; Zhou, 2019). Figure 2 shows a visual layout of Steven's (1946) levels of measurement (further resources are available here)[4] (Bond & Fox, 2015; Coaley, 2010; Michell, 2008).

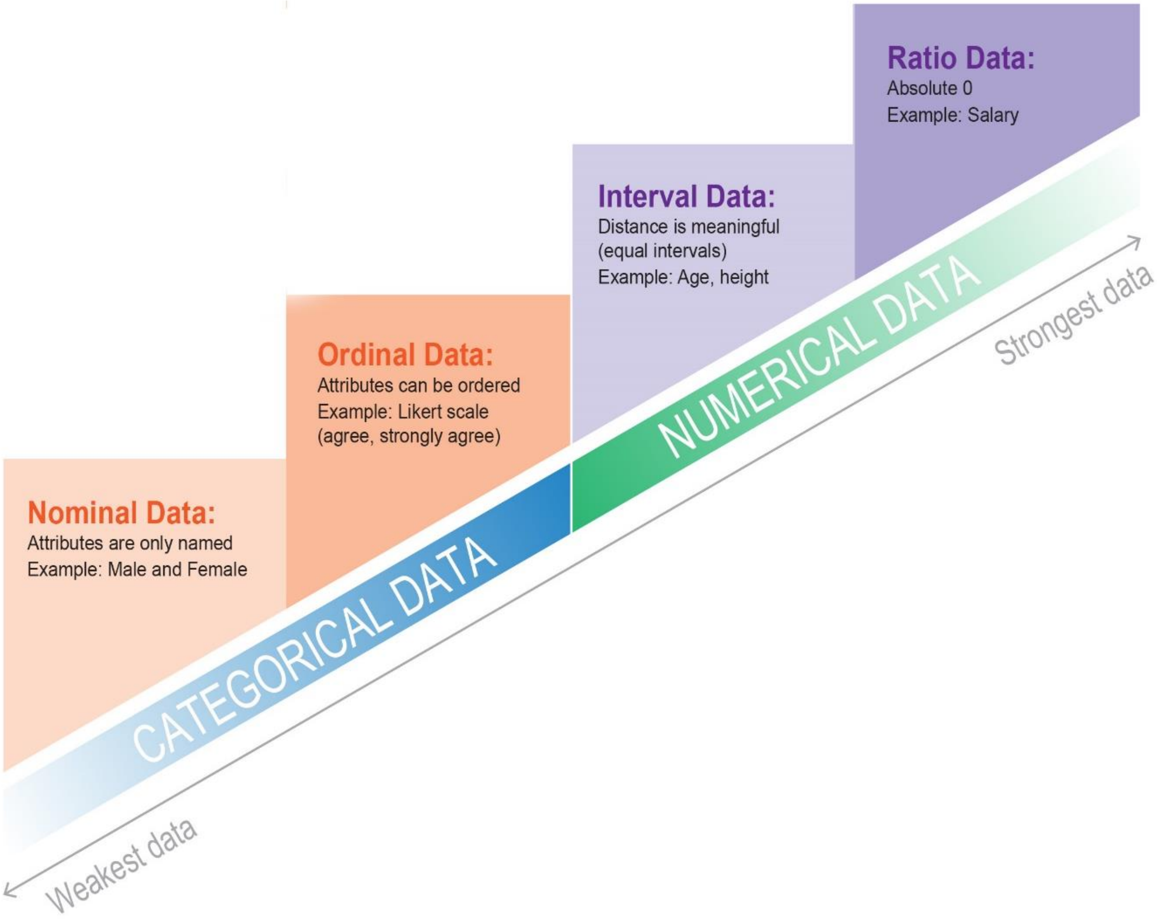
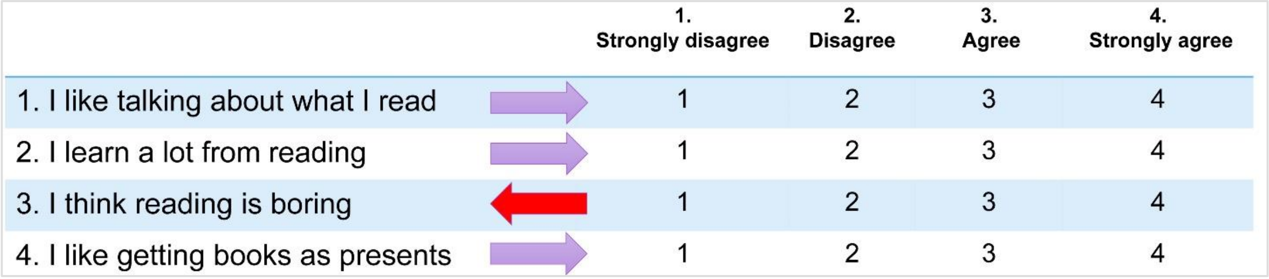


Figure 2. Levels of measurement (Source: Author)

[4] Data Science & Statistics: Levels of measurement video at
https://www.youtube.com/watch?v=eghn__C7JLQ
Types of Data: Nominal, Ordinal, Interval/Ratio - Statistics Help at
https://www.youtube.com/watch?v=hZxnzfnt5v8

The researcher should be able to classify their variables according to the types of data evident in Figure 2. Researchers tend to treat summative scores as interval, for example, the total score obtained on an anxiety index. But raw scores without modelling are ordinal and lack equal intervals (Wright, 1993). What about the percentage a child scores on a mathematics test; is that ordinal or interval? From a Rasch perspective, the summative score would be considered interval data if there is evidence of objective measurement. Objective measurement is achieved if the data and items fit the Rasch model and adhere adequately to the measurement assumptions.

The assessment designer should know their instrument intimately, including the composition of the scales, the rating categories, and any subscales or separate constructs in the questionnaire (for multi-dimensional instruments). Constructs should be well operationalised. Types of items should be defined in the assessment framework, as demonstrated in the Appendix. Any items that are negatively phrased or designed to show the opposite of the underlying construct should be re-coded during the data preparation phase. In Figure 3 below, the construct is *the enjoyment of reading*. The more a child agrees with a statement, the more they should enjoy reading. However, question three is phrased opposite to the underlying construct.



		1. Strongly disagree	2. Disagree	3. Agree	4. Strongly agree
1. I like talking about what I read	→	1	2	3	4
2. I learn a lot from reading	→	1	2	3	4
3. I think reading is boring	←	1	2	3	4
4. I like getting books as presents	→	1	2	3	4

Figure 3. Example of an item that requires reverse scoring (Source: Author)

Before analysis commences, item three should be reverse-scored so that 1 = 4; 2 = 3; 3 = 2 and 4 = 1. Rescoring can easily be calculated in SPSS, Excel or similar statistical software.

THE FASHIONS OF PSYCHOMETRICS

Measurement in psychology has a long and controversial history, for both its scientific legitimacy as well as its uses and misuses (Bradley & Brand, 2016; Bruschi, 2018; Corcoran, 2014). Classifying psychology as science requires accurate and useful measurement, and it is here that Rasch theory and Item Response Theory (IRT) emerged as applications. Before Rasch and IRT, Classical Test Theory (CTT), also called traditional test theory, served as the default. CTT contains the foundations of modern psychometric theory, including the use of reliability coefficients, accounting for measurement error and item discrimination. The limitations of CTT are well documented, and many are easily recognisable (Boone, Staver, & Yale, 2014). The most often cited limitation of CTT is sample dependence for reliability coefficients, item difficulty and item discrimination (Rusch, Lowry, Mair, & Treiblmaier, 2017). CTT does not offer invariant measurement because it does not separate person and item parameters. When there is evidence of invariant measurement, item difficulty and hierarchy remain consistent when applied to comparable samples. Rasch measurement offers tests of invariance as well as modelling these through parameter separation. CTT assumes a linear relationship between raw scores and the latent trait, an assumption that may be erroneous. Rasch theory tests the assumptions by constructing linearity through separation parameters (Harwell, Gatti, & Linacre, 2002). CTT assumes an interval scale which does not hold for the categorical nature of items. Total scores are not interval if the distance between items and categories are not equal and ordered. Rasch tests assumptions of order, the dimensionality of the underlying construct, and invariance across groups and creates equal-interval scales (Franchignoni et al., 2012). The CTT assumption that error and true scores are uncorrelated is also problematic as distortion happens at the extreme ends of the scale and error influences derived scores. Distortion of error means no equal intervals and bias in our measurement (Allen & Yen, 2002; Massof, 2011). When we accurately model the measurement error and account for it with Rasch theory, we obtain an interval scale and reduced bias. Sophisticated applications, such as Computer Adaptive Testing (CAT), cannot be done with CTT. For psychology and other social sciences to attain accurate and useful measurement, Rasch or IRT models are recommended for instruments design, refinement and currency. Ordinal data should not be handled as though they are interval, as the conclusions drawn from parametric statistics would then be misleading. Adherence to measurement principles should be assessed and confirmed, or scores should be converted to an interval scale, both of which can be done by Rasch theory if researchers want to use parametric statistics to draw conclusions from their assessments and compare groups fairly. Despite the limitations of CTT mentioned here, it is important to note that the user need not be confined to an either-or choice. CTT, Rasch and IRT can be viewed as complementary methods (Gerriet, Valerie & Jonathan, 2014). CTT limitations do not mean the theory does not offer useful information about assessments (Nolte, Coon, Hudgens, & Verdam, 2019). CTT results can be used to improve and refine items when the limitations of the theory are taken into consideration. CTT can also be combined with Rasch or IRT for more comprehensive results. For example, the use of structural equation modelling (SEM) to model groups or construct loadings via factor analysis (CTT) can be achieved with Rasch to model item and person parameters. The combined use of CTT and Rasch can create a more comprehensive model of instrument functioning (Krägeloh et al., 2013). A pragmatic approach is encouraged where the designer and user of assessments apply multiple tools to assure high quality, useful measurement (Boateng, Neilands, Frongillo, Melgar-Quiñonez, & Young, 2018).

Rasch, Item Response Theory or Structural Equation Modelling?

Rasch theory parametrises item difficulty (Andrich & Marais, 2019). In Rasch theory, guessing is seen as an aspect of ability. Persons of lower ability are the most likely to guess the answer to a multiple-choice question. Item Response Theory (IRT) can include four[5] parameters and accounts explicitly for guessing. The concepts of difficulty, discrimination and pseudo-guessing are explained in the jargon section. Below is a summary of what the one, two and three-parameter models in IRT include:

1. Difficulty (location), also called the one-parameter model (1PL) or Rasch model.
2. Difficulty (β) and Discrimination/Slope (α): also called the two-parameter model (2PL).

[5] Note that a four-parameter (4PL) model is available, wherein anomalous responses, such as carelessness, is parameterised.

3. Difficulty (β), Discrimination (α) and the Lower Asymptote (γ). The lower asymptote is also known as the pseudo-guessing parameter. The model is designed for assessments with Multiple Choice Questions (MCQs) and is called the three-parameter logistic model (3PL).

An essential question you as a researcher must ask is: *what drives the analysis? Is it theory or data?* In Rasch philosophy, data should conform to the Rasch model because the model is designed to test the principles of measurement (Boone & Rogan, 2005). This is in contrast to IRT, where the best fitting model for the data is sought so that the data drives the analysis. The Danish mathematician, Georg Rasch, was a proponent of objective measurement, and he argued that the model should indicate the quality of the data (Rasch, 1992; 1980). IRT takes an exploratory approach, where the data leads to the analysis. The IRT user adjusts the model until the best fit is derived for the data. The user would choose the Rasch model rather than the two or three-parameter models when the focus is on creating a useful measurement rather than finding a best-fitting model for the data (Bond & Fox, 2015). From a Rasch perspective, IRT violates sample independence when parameters are fixed to an arbitrary value and allowed to cross (Shaw, 1991; Wright, 1992). The sample independent nature of Rasch modelling is one of its biggest advantages, providing evidence for invariant measurement and generalisability of findings (Smith, Wakely, de Kuif, & Swartz, 2003). The Rasch model strictly measures whether the data fits the model, whereas other IRT models try to fit the model to the data to obtain the best fit and less biased parameter estimates. Rasch users consider their models to be separate from IRT, due to the philosophical differences and strict assumptions of the Rasch model. IRT users tend to view Rasch as an application of the one-parameter model and thus within the range of IRT uses. The one-parameter model and the Rasch dichotomous model are the same mathematically. Interested readers can consult further resources [6] . Andrich (2004) supports a Rasch perspective and Hambleton, Swaminathan and Rogers (1991) provide the IRT point of view.

[6] Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? Medical Care, 42(1), 16.

Another approach gaining popularity is the use of Structural Equation Modelling (SEM) to assess instrument functioning and measurement invariance. The use of SEMs to assess invariance could be used in conjunction with Rasch or IRT as complementary tools (Cohen & Swerdlik, 2018) or independently to evaluate the data. SEM methods are not discussed in the current chapter as knowledge of Rasch theory is sufficient for assessing instrument functioning, especially for novices. Rasch and IRT models are mostly unidimensional models, but multidimensional models are available and growing in popularity for Rasch, IRT and SEMs. I recommend that readers gain proficiency in applying unidimensional models, and then try multidimensional models if the models are relevant to the research being conducted. The present chapter takes a pragmatic approach, one in which all statistical models are seen as tools. The responsible scientist finds the tools that are the most appropriate for the analysis and provides accurate answers. The chapter guides the reader in the application and interpretation of Rasch theory. The theory is considered the most practical path to objective measurement, and the model is not adjusted to fit the data. Instead, the instrument items must be adjusted to improve measurement, because the measurement is defined *a priori*. For psychology to be a science, measurement needs to have the same rigorous standards as required by the natural sciences.

Internal and external validity

Rasch and IRT models assess the internal reliability and validity of inferences derived from instruments. In the design and use of any instrument, the internal functioning of the items and tool is the first and most crucial step. Rasch statistics provide a set of prescriptive criteria for checking the degree to which measurement has been successful. Pilot studies to gauge the internal functioning of an instrument should include a trial of the instrument with a sufficient sample size and target population. The sample size depends on the number of parameters to be estimated, the specific Rasch model utilised and the required stability of

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage Publications.

parameters. Some authors offer rules of thumb; for example, a five-point Likert scale should have 10 – 20 participants per category, a sample of +/- 100 participants (Chen et al., 2014). The researcher should consider whether any particular rule of thumb is practical for their study and calculate their own required sample size. Even small samples (>30) can offer useful information if there are fewer parameters to estimate, for example, the dichotomous model, or if the instrument has been through more than one round of piloting. The Rasch perspective of validity emphasises construct validity. Construct validity aims to measure the operationalised construct so that it is neither over nor under-represented (Messick, 1989), but instead the items offer a comprehensive, useful indication of the construct's functioning within the intended population.

Only the quantitative aspect of evaluating the instrument is discussed in the current chapter. Qualitative analysis for design and refinement of instruments is both recommended and could be crucial for understanding the statistics; for example misfitting items may have phrasing problems that can be identified through discussion with respondents (Recabarren, Mallinckrodt, & Miles, 2016). Qualitative analysis can include evaluations by subject matter experts (SMEs), interviews or focus groups with participants, written comments on the instrument and think-aloud methods (Power, Lemay, & Cooke, 2017). Once sufficient evidence has been garnered to demonstrate that the instrument is functioning well internally, the external validity can be investigated (for example, concurrent validity). Take note that if there are internal problems with a questionnaire or test, it may negatively affect external validity. Internal and external reliability and validity are intertwined.

RASCH MEASUREMENT THEORY

Rasch analysis is a method for obtaining objective, fundamental, additive measures (Linacre, 2019, p. 35).

Rasch models have been designed to accommodate a range of item types. The dichotomous Rasch model simply states that the probability of the event (answering correctly/agreeing) is a mathematical function of the difference between item difficulty (or endorsability) and ability (or endorsability) as demonstrated in Figure 4.

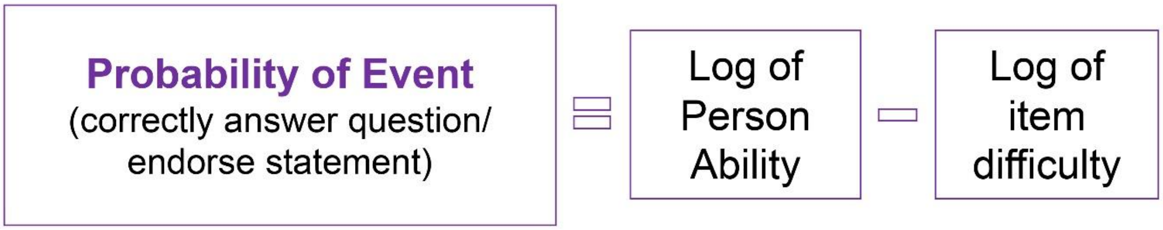


Figure 4. Rasch dichotomous model representation (Source: Author)

The difficulty of each item is based on how many people answered it correctly/agreed compared to the overall ability/agreeability of all the people who answered. A probability of correctly answering/agreeing is calculated for items and persons independently. For example, we ask people in a questionnaire if they disagree (0) or agree (1) with the statement: "I feel anxious most days". The item has a difficulty of 1.12, then a person who also has an ability of 1.12 has a 50% chance of agreeing with the statement. A person with an ability of 2.30 has a 76% chance of agreeing with the statement. The probability can easily be calculated by substituting the values of item difficulty and person ability into the formula: Probability = (exponential (person ability – item difficulty)) / 1+ (exponential (person ability – item difficulty)).

Rasch Models

The most commonly used Rasch models are the Dichotomous and Polytomous models. The polytomous model is also known as the Partial-Credit Model. The type of model applied depends on the item. The Rasch dichotomous model was designed for items that have only two options, for example:

- Right or Wrong (0 or 1)
- Yes or No
- Disagree or agree

The polytomous model is applied to items with more than three options, as is the case with Likert type items, or test items where a person can score out of two or more. Item difficulty for polytomous items is calculated as the intersection of the lowest and highest category, with difficulty estimates given at all the threshold intersections. Instruments can contain a mixture of items with different categories when the Partial-Credit Model is applied. Rasch software can handle items with various options by analysing each item on its own scale. The software applies the appropriate model.

The Rasch Logit Scale

The Rasch model converts the item and person estimates to the natural logarithm, a conversion which results in an interval scale (Wright & Stone, 1999). The logit scale ranges from negative infinity to positive infinity. But most logit scales fall between -5 to +5 on the scale, and scales beyond this range may indicate a heterogeneous population with a range of abilities too wide to measure with the intended instrument. Rasch scales are set to a mean of 0 and a standard deviation of 1. The logit scale may be difficult for people to interpret. Most software packages have options to rescale logits to more familiar numerical ranges, for example, a scale of 0 to 100. Syntax can be used to rescale item or person logits scores derived from the software with the SPSS syntax shown below. To use the syntax, calculate the logit mean and standard deviation and insert them into the following formula:

Compute Logits_Rescaled = 50 + 15 * ((Logits_Variable – Mean)/StandardDeviation).exe.

To rescale the data in SPSS, remember to export the person or item measures and add them as a variable. In the example above the exported logits has the variable name "Logits_Variable". The mean of 50 and standard deviation of 15 shown in the example can be changed to a range to suit the user.

Assumptions of Rasch

The premises of Rasch models are assessed with the statistics provided by the software. Failure to meet assumptions could indicate measurement problems (Iramaneerat, Smith, & Smith, 2008). When designing an instrument or choosing to use an existing one, the underlying construct should be operationalised. The aim is to measure an aspect of human behaviour, thought or cognition with a set of items. Rasch theory places items in order of difficulty, and respondents (persons) in order of their ability/endorsability and then aligns persons and items on the same scale (Fox & Jones, 1998). Rasch theory provides empirical evidence of construct validity. The assumptions of the models are summarised below (Bond & Fox, 2015; Boone et al., 2014; Iramaneerat et al., 2008):

1. **Internal reliability and validity** obtained respectively through consistency of answers and coherency of the construct. Rasch theory assesses the reproducibility of items through reliability indices. Construct-representation is examined via fit statistics and targeting indices, such as item-person maps (Baghaei, 2008).
2. **Equal intervals** between units of measurement regardless of position on the scale, i.e. each question or category are equally more difficult/endorsable than the previous one.
3. **Additivity and concatenation:** the items are positively related to the underlying construct, and each item adds to more of the construct in an ordered way.
4. **The unidimensionality** of the construct/latent trait. Unidimensionality is a fundamental concept in scientific measurement that assumes that one attribute of an object is measured at a time. If there are theoretical sub-constructs in the instrument, they should not be strong enough to compete with the central construct. If sub-constructs are strong enough to form a factor, those items would be seen as creating a different construct to the main one and would have to be analysed separately.
5. **Invariance,** the construct as measured by item estimates remain constant over time, comparable groups and places. Invariance is strongly linked to reliability and assessed with statistics such as differential item functioning.
6. **Order**: there is a logical progression in obtaining more of the construct. For example, crawl, walk and then run. Persons who are more able or agree more have a higher likelihood of correctly answering items or agreeing with difficult statements
7. **Categorical Order**: Categories should increase monotonically with more of the underlying trait when the scale is designed for higher scores to indicate more of the

construct. Categories are orderly when respondents can distinguish between options offered, and the options make sense to respondents when they apply them to items.

8. **Local independence of items:** When items are highly correlated, they may be redundant or repetitive and cause noise in the data. But if the highly correlated items represent different aspects of the construct, do not remove the items as this may lead to construct under-representation.

Rasch theory and decolonisation

The link between a measurement model and the call for decolonising psychology may not seem evident at first glance. Nevertheless, what we measure and how we measure guides scientific endeavours. The apartheid government utilised an unethical and methodologically flawed approach to measurement to justify inhumane practices (see Bedell, van Eeden & van Staden, 1999; Milner, Thatcher & Donald, 2014; Sehlapelo & Terre Blanche, 1996). In contrast, Rasch models are grounded in theory which emphasises measurement invariance and that instruments should be fair and robust so that constructs are measured in the same way for diverse groups at different periods (Engelhard, 2008; Molenaar, & Borsboom, 2013). Rasch theory recommends the refinement or reconceptualization of items and constructs that violate measurement invariance. Constructs are not facts but are rather born through human lenses through which we organise behaviours, ideas and opinions into categories. Rasch theory is the mathematical evaluation of construct validity, a test of whether items and instruments create useful tools. Rasch theory provides researchers with clear indicators of whether instruments are fair and useful. Designing and using instruments enforces the perception that constructs are credible. Rasch models challenge assumptions of construct validity and encourage the user to think critically about items, instruments and constructs. Only if the items and instruments suitably meet the assumptions of Rasch theory is there evidence that invariant inferences can be derived from the test or questionnaire in the context in which it is being utilised. Rasch theory also directs the revision of instruments. When items are potentially biased towards one group, as may be found when differential item functioning is examined, Rasch theory suggests that the reason for the bias be identified and items improved. Rasch theory can be applied to decolonise psychological assessment by gauging the validity of western constructs for an African context. When constructs and instruments designed for Eurocentric milieu do not function well in our local context, African researchers should take up the challenge to do the extreme knowledge work needed to create locally relevant paradigms. Decolonisation is a process which requires measurement designed for local linguistic and cultural contexts through the rigorous application of scientific principles.

Guide to Rasch Analysis

There is no single recipe to use when conducting a Rasch analysis. The statistics required depend on whether a new instrument is being designed, an existing assessment is being refined or an item bank is being built. Regardless of the purposes, the following aspects are recommended when conducting the analysis:

- Overall instrument and item functioning (reliability, fit statistics, global model fit)
- Unidimensionality of underlying construct
- Local independence of items
- Category and threshold functioning
- Differential item functioning (DIF)
- Person and item alignment

The software

There are various Rasch software packages available, both paid and freeware. Freeware packages are available, but the user may need skills, such as writing syntax, to utilise them. The freeware, R, has a package (Mirt) which offers most IRT models in use today, is a state-of-the-art application, and is generally recommended. There is value in learning to use R if one is going into the IRT domain. The two packages presented here are both paid software with more user-friendly GUI interfaces. Winsteps is less expensive[7] and has more online resources available (Linacre, 2019). The Winsteps licence offers access for life and sends the user access to updated versions as they become available. Rasch Unidimensional

[7] In 2020 priced as $149 (US dollars)

Measurement Models (RUMM) software is more expensive but has more elegant figures and presents statistical significance in a more user-friendly format (Andrich, Lyne, Sheridan, & Luo, 2010). When buying a version of RUMM, the user has access to that version for life but no option for automatic upgrades to newer versions. Each software package has its own pros and cons. Both Winsteps and RUMM can be used to conduct a useful Rasch analysis. Outputs from different software packages will vary, not surprisingly considering the different types of likelihood analysis applied, but the differences tend to be minor (Robinson et al., 2019). More information on non-commercial software options is available[8]. The user is encouraged to explore different options and preferences for software depending on his/her needs. Researchers and students who would like to have access to Winsteps are encouraged to contact the author to arrange a sponsored licence.

Preparing the data

Before data can be imported into Winsteps or RUMM, it should be prepared in either Excel or SPSS. For the current paper's example, preparation in SPSS and Excel are shown. Setting up a codebook and an assessment framework is strongly recommended. Record the scale of each item as well as how values were coded, including codes used for missing data (see Appendix for example). Two variables are essential when applying the Rasch measurement model, the unique identification number (ID) per person as well as the questions captured at the item level. If different groups were assessed, for example, men and woman, you may want to add person factors. The three types of variables you need to set up your data are:

- **Unique identification number:** each person should have a unique number which identifies them. In SPSS, check that there are no duplicate cases by going to data, and selecting Identify Duplicate Cases. In excel, click on Conditional Formatting, Highlight Cell Rules and then select Duplicate Values. The presence of duplicate values may indicate data entry problems and cleaning may be required.

[8] https://www.rasch.org/software.htm

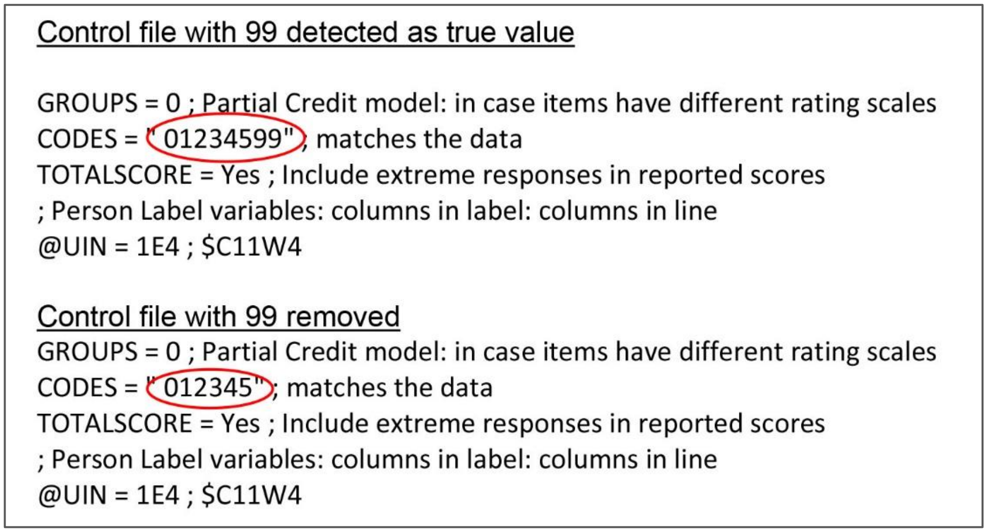
- **Person factor(s):** Any demographic or group identifying variables. Measurement invariance means that the items function the same way for groups who have a similar ability. Adding group identifying variables helps you to assess measurement invariance between groups.
- **Items:** The questionnaire items/test questions captured at the item level.

Data importation

Importing data into Winsteps or RUMM generally follows the same steps. Where the two programmes diverge, this is indicated.

STEP 1: Clean and recode

- **Data cleaning or screening:** Generate descriptive statistics of person factors and items. Check that all variables are in the appropriate numerical range.
- **Recode variables** if and where required, especially if items have been designed to negatively correlate to the underlying construct, in which case the items must be reverse scored.
- **Missing** data should have one value to indicate that a question was not answered. For RUMM, the code to show missing data should have the same number of characters as the variable. The value "9" could be used to indicate missing values if the total score of the item does not exceed 8. In Winsteps, any or multiple values could be used as missing, but the user would have to specify this in the control file. The easiest option for the Winsteps user is to replace the missing value with empty values before importation. The example in Figure 5 below shows a control file where the value for missing, "99", was detected as a true value for the items. The user would have to delete the 99, as shown in the example below.



Control file with 99 detected as true value

```
GROUPS = 0 ; Partial Credit model: in case items have different rating scales
CODES = "01234599"; matches the data
TOTALSCORE = Yes ; Include extreme responses in reported scores
; Person Label variables: columns in label: columns in line
@UIN = 1E4 ; $C11W4
```

Control file with 99 removed

```
GROUPS = 0 ; Partial Credit model: in case items have different rating scales
CODES = "012345"; matches the data
TOTALSCORE = Yes ; Include extreme responses in reported scores
; Person Label variables: columns in label: columns in line
@UIN = 1E4 ; $C11W4
```

Figure 5. Winsteps control file with missing data value removed (Source: Author)

STEP 2: Reorganise items

- **Reorganise items according to type and maximum score.** This step is necessary for RUMM and advisable for Winsteps.
- **Move questions (variables)** in the excel data sheet so that all multiple-choice questions are together. All questions scored out of one mark should be grouped together; all questions scored out of two marks together and so forth. Likert type items should be grouped together to have the same number of categories (options).
- **Organise multiple-choice questions (MCQs)** together according to the key (answer), for example, grouping all the variables with the key A, B, C and then all the Ds.
- **Group constructed response questions (CRQs)** from the smallest to the highest number of categories (1, 2, and 3).
- If all questions have the same number of categories, for example, all items have the same 5-point Likert scale, then no reorganisation of items are necessary.
- The data sheet should have the unique identification number first, followed by any person factors and then the items arranged as recommended above. Make sure the data file contains only the items needed for the Rasch analysis.

STEP 3: RUMM final preparation in excel

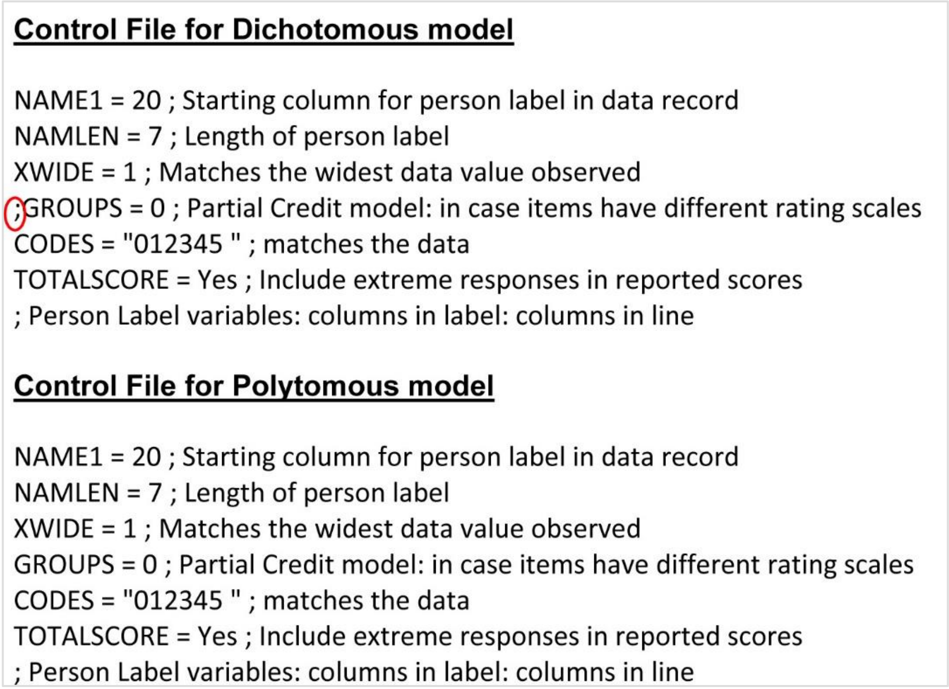
The following steps are only required for entering data into RUMM:

- **Set column width for use in RUMM**. All columns should be set to the maximum number of characters. For example, if you have a sample of a thousand participants, their identification numbers will go up to 1000. Therefore, the ID variable column width would have to be set to 4. In contrast, other variables, such as questionnaire items, may only need to be set to 1 if that is the maximum number of characters.
- **Delete the variable names**: Again, this is only necessary for RUMM. Delete the first line of the excel sheet which contains the variable names.
- **Save the file as Formatted Text (Space Delimited), aka PRN extension:** This is the format required for RUMM. Winsteps can read the excel file as long as you close the file before opening it in Winsteps. For Winsteps, make sure the sheet containing the data is the first one in the file before saving it.

STEP 4: Import data into Winsteps or RUMM

- Open the Winsteps or RUMM software[9]
- **Winsteps**: Special note for Winsteps: To apply the Rasch Polytomous (Partial Credit) model, you must remove the semi-colon in the control file (Linacre, 2012). Not removing the semi-colon indicates that all items are on the same scale, and the Dichotomous or Rasch Rating Scale model will be applied. When you want to use the Partial Credit Model, delete the semi-colon before the *groups command*, as shown in the example in Figure 6. This is a distinctive aspect of the software.

[9] http://www.winsteps.com/a/winsteps-tutorial-1.pdf



Control File for Dichotomous model

```
NAME1 = 20 ; Starting column for person label in data record
NAMLEN = 7 ; Length of person label
XWIDE = 1 ; Matches the widest data value observed
;GROUPS = 0 ; Partial Credit model: in case items have different rating scales
CODES = "012345 " ; matches the data
TOTALSCORE = Yes ; Include extreme responses in reported scores
; Person Label variables: columns in label: columns in line
```

Control File for Polytomous model

```
NAME1 = 20 ; Starting column for person label in data record
NAMLEN = 7 ; Length of person label
XWIDE = 1 ; Matches the widest data value observed
GROUPS = 0 ; Partial Credit model: in case items have different rating scales
CODES = "012345 " ; matches the data
TOTALSCORE = Yes ; Include extreme responses in reported scores
; Person Label variables: columns in label: columns in line
```

Figure 6. Example of removing semi-colon to set the control file to polytomous data (Source: Author)

- **RUMM**: Detailed instructions[10] can be found in the user manuals installed with the software, located in the C drive in the installation folder (RUMM Laboratory, 2015).

Analysing and interpreting Rasch results

Rasch statistics should be understood holistically; all evidence should be examined as a whole and interpreted through the lens of what is known and expected of the construct. The old adage: *don't let the numbers do the thinking for you* is especially true of Rasch philosophy. The aim is not to design items or assessments that fit the model perfectly or even non-significantly. The instrument and its items should be *useful* for measuring the intended samples. Pragmatism is the order of the day, and dogmatic statistical decisions should be avoided (Linacre, 2019). Table 1 shows the most frequently used Rasch statistics for

[10] http://www.rummlab.com.au/

Winsteps (Linacre, 2020) and RUMM (RUMM Laboratory, 2015), respectively, with guidelines on how to interpret the results (Bond & Fox, 2015; Boone et al., 2014; Jafari et al., 2012; Marais & Andrich, 2007). The location of the statistics in the software is presented first, followed by the interpretation. For example, to find the global fit statistics in Winsteps, consult Table 44.1 in the software.

Table 1. Interpretation of Rasch statistics produced by Winsteps and RUMM

Measurement Criteria	Winsteps	RUMM
Data fit Rasch model (Global fit statistics)	**Table 44.1** **Log-likelihood chi-square:** A small $\chi 2$ which is non-significant (p>0.05) is desirable and indicates the data fit the Rasch model **Global Root-Mean-Square Residual (RSMR):** An expected value is shown. If the real value is smaller than the predicted value, this indicates a better fit. **Capped Binomial Deviance:** values smaller than the expected = better fit	**Test-of-Fit Details: Summary Statistics** **Total item-chi square ($\chi 2$):** A small $\chi 2$, which is non-significant (p>0.05) is desirable and indicates the data fit the Rasch model **Power of Analysis of Fit rating:** A rating which provides an interpretation for the person reliability score of the test, rating it from too low up to excellent.
Reliability statistics	**Table 3.1 Summary Statistics** Person reliability is similar to Cronbach's Alpha and indicates reproducibility. Suggested interpretation: <.50 = unacceptable >.70 = good >.80 = very good	**Test-of-Fit Details: Summary Statistics** Coefficient Alpha (α) provided when you run data without missing. Suggested interpretation: <.50 = unacceptable >.70 = good >.80 = very good

Measurement Criteria	Winsteps	RUMM
	>.90 = excellent, 3 - 4 levels of ability/agreeability can be distinguished **Person-separation index:** below 2 = weak, indicates items may not be sensitive enough to distinguish high achievement/agreement from low. **Item reliability interpretation:** At least >.70 for acceptable item reliability. Low reliability may indicate sample was not big enough to locate item order on the latent construct (meaning for construct) **Item separation index:** below 3 = weak, indicates sample may be too homogenous or too small to establish item difficulty hierarchy **Root Mean Square Error of Approximation (RMSEA):** Mean of measurement standard error. 0.05 or smaller is good fit, between 0.06 and 0.08 is reasonable and > 0.1 = poor fit	>.90 = excellent **Person-Separation Index (PSI) interpretation:** >0.7 minimum accepted level and indicates statistically differentiable two groups >0.9 = 4 ability groups are distinguishable **Fit Residual Item Mean**: The ideal is the mean, which is set to 0. Good range: -0.50 to +0.50 **Standard Deviation *(SD)*:** greater than ½ an *SD* may indicate problems, i.e. > 0.50
Item functioning	**Table 10 Item Fit Order** Infit Mean Square (MNSQ) – inliers	**Test-of-Fit Details: Individual-Item-Fit** **Fit residual value guidelines:**

Measurement Criteria	Winsteps	RUMM
	Outfit Mean Square (MNSQ) – outliers **MNSQ general guidelines**, but high stakes testing may require stricter criteria: 0.5 to 1.5 good range <0.5 useful but duplicative 1.5 to 2.0 useful but noisy > 2.0 unexpected responses, investigate **ZSTD:** Ranges from -2 to +2. Values outside the range considered suspicious. Examine MNSQ first, look at outfit ZSTD if MNSQ indicates a problem	- 2.5 to +2.5 ideal range Probability (*p* smaller than Bonferroni adjustment) indicates misfit is statistically significant. Bonferroni adjustment is applied automatically. Items with large fit residuals AND which are statistically significant should be investigated. **Discrimination:** If an item is too easy, the residual value will be negative, and when an item is too difficult, it will be positive. Equivalent to the Winsteps overfit and underfit terminology.
Unidimensionality	**Table 23** **Principal Component Analysis (PCA):** Eigenvalues > 2 indicate items could be forming another dimension. Examine items that cluster together and consider whether they represent a separate construct. If they do, analyse those items separately.	**Test-of-Fit Details: Residual Principal Components** **Principal Component Analysis (PCA):** Eigenvalues > 2 indicate items could be forming another dimension. The summary shows the percentage of total variance accounted for by each factor once Rasch variance has been removed. Potential multidimensionality should be investigated.

Measurement Criteria	Winsteps	RUMM
Local independence of items	**Table 23.99**[11] **Pearson correlation coefficient (*r*):** Highly correlated items may violate the assumption of local independence. Investigate items where: $r > .70$	**Test-of-Fit Details: Residual correlations** **Residual Correlation Matrix:** Highlight values above 0.3 (rule of thumb) to identify items which may unduly influence one another and be overly correlated
Category and threshold functioning	**Graphs: Category probability curves** Categories should be ordered and increase monotonically. A rule of thumb is at least ten observations per category or at least 5% of responses per category. Categories which no-one or very few people chose/correctly answered are problematic and could distort data. Disordered thresholds should be investigated and are only a threat when narrow intervals are undesirable. Check ISFILE: MNSQ greater than 1.3 and unweighted % smaller than 10% indicate DIF is significant	**Item characteristics: Category Probability Curves and Other Outputs: Threshold Map** Categories should be ordered and increase monotonically. Item locations should be close to the mean, and the chi-square should be non-significant Also consult: **Item Characteristics Curves** where person means in each class interval should be close to the theoretical curve.

[11] Watch for more details: https://www.youtube.com/watch?v=sna19QemE50

Measurement Criteria	Winsteps	RUMM
Differential item functioning	**Table 30 Item DIF** **Two groups consult Table 30.1 for pairwise comparisons** In the table look at the DIF contrasts > 0.5 and Rasch-Welch probability and/ or Mantel probability smaller than 0.05. Items with large contrasts which are also statistically significant (*p*<0.05) should be investigated. **More than two groups consult Table 30.2 or Table 30.3 for multiple comparisons.** **Table 30.4 provides indication for non-uniform DIF.**	**Item characteristics: Item Characteristics Curves** Click on *DIF Summary* and select *Highlight probs*. The items which have statistically significant DIF between groups are highlighted in red and should be investigated. Bonferroni corrections are applied automatically **The set of columns indicate non-uniform DIF**
Targeting: Person and item alignment	**Table 12 Item Wright Map** Examine the map for items and persons alignment. Are there gaps? Gaps could indicate aspects of the construct not covered or a lack of items to target specific groups. Is the mean of the persons and items far apart? When items and person align well, there is evidence of good targeting.	**Further Outputs: Item Map and Person-Item Distribution** Look at the maps provided for alignment between persons and items. When items and person align well, there is evidence of proper targeting. The Person-Item Distribution map in RUMM also allows you to examine the spread between different groups in your data if you included person factors.

The user is encouraged to look at all the evidence and evaluate the instrument in conjunction with other sources. Removing an item should be a last resort. Consider first how the item contributes to the construct, and if removing the item could cause construct under-representation. Try to improve the item, by for example, rephrasing/retranslating or improving distractors. Improving the fit of individual items is more important than the global fit. Better fitting items will also lead to a better overall model fit. The chi-square is an indication of how well the data fit the Rasch measurement model; a significant χ2 indicates that the data and the model are significantly different from one another. However, this misfit should not be interpreted from an absolutist perspective; instead, it is one of many indicators to consider. The chi-square test is very sensitive to sample size, and large samples may give a significant result. When sample sizes are larger than 200, interpret the Root-Mean Square Residual, which is available in Winsteps (Tennent & Pallant, 2012).

Unidimensionality is a complex aspect of test development, and you should examine both the Principal Component Analysis (PCA) of the standardised residuals as well as the local independence of the items. The PCA results in both Winsteps and RUMM show the variance after removing the Rasch construct (item difficulty). In Winsteps, the contrast groups can be examined to see if the disattenuated correlations form sub-dimensions. In RUMM, the principal components can be investigated on the PC loadings tab where the user can select the “highlight above” option. When evidence of multidimensionality emerges, look at the items causing the potential problems. Are those items indeed forming a separate construct? Examine whether or not the test is unidimensional *enough* to justify treating all the items as forming one construct. Investigate the local independence of items to make sure items are not redundant. Consult subject matter experts and literature to understand why items are overly correlated and retain or remove items based on the theoretical representation of the construct.

Both Winsteps and RUMM provide category probability curves to examine how well the categories function. Designing categories that are sensible for respondents can be challenging, and social desirability responding is a common problem. Agreement scales may lead to respondents agreeing or strongly agreeing with most or all statements, and the designer is encouraged to explore other options such as asking about the frequency of behaviours or experiences, asking participants to rate indirect statements, and asking participants about concrete aspects of the construct. In Figure 7, an example is shown regarding how to interpret a category probability curve graph.

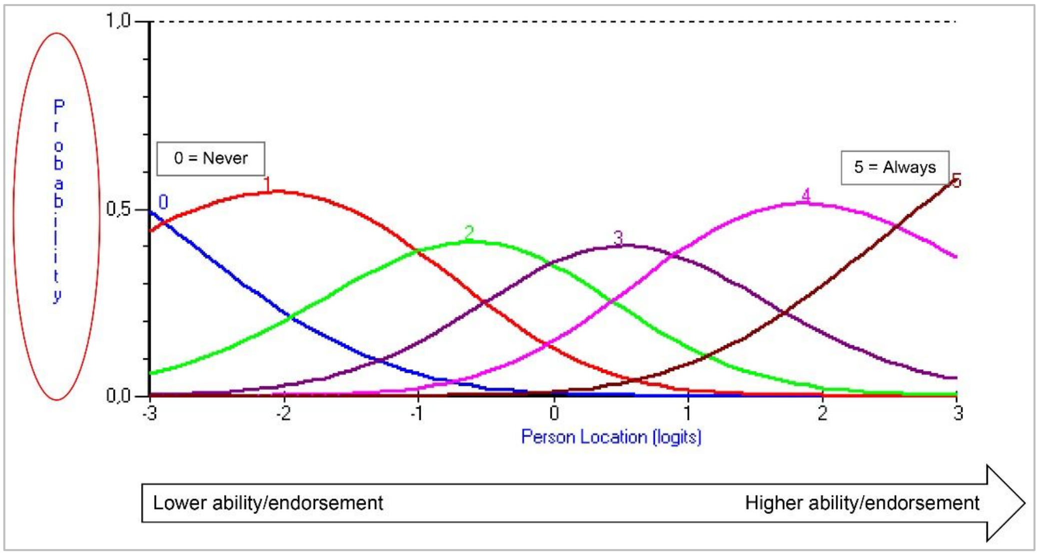


Figure 7. How to interpret a category probability graph (Source: Author)

On the vertical axis, the probability of a category being chosen is shown, and on the horizontal axis, the person location (ability/agreement) is shown. Each curve represents a category; for example, the blue curve is the option of "never", coded as 0, and the magenta colour shows the option of "always", coded as red. On the graph, we want to see all options having a probability of being endorsed, and that probability should increase with each subsequent option and more person agreement/ability. So, the "always" option should be most probable for persons at the highest end of the logit scale, as is evident in this graph.

When examining differential item functioning to assess measurement invariance, keep in mind that there are two types of DIF:

- **Uniform DIF** is when one group has a consistently different probability of agreeing/answering correctly, despite having the same underlying person location as another group. For example, children answering the French version of an item are significantly less likely to get the correct answer even if they have the same underlying ability children answering the English version have, due to a translation problem.
- **Non-Uniform DIF** is when one group has an inconsistent probability of answering correctly/endorsing an item compared to another group.

Empirical Example

South Africa participated in an international assessment project where teachers in the Western Cape Province rated Grade 1 children on an Attention-deficit/hyperactivity disorder (ADHD) questionnaire. In total, 1 572 children were evaluated, and the sample was randomly selected per class and stratified on gender to obtain a balanced sample (Mtsatse & Combrinck, 2018; Tymms, Howie, Merrell, Combrinck & Copping, 2017). The instrument is an existing rating scale which has been used in the United Kingdom, Brazil, China and other countries (Merrell, Sayal, Tymms, & Kasim, 2017; Merrell & Tymms, 2005). However, this is the first time the instrument was used in South Africa, and the internal validity and reliability were assessed by applying the Rasch measurement model. The instrument has 18 items which are based on the Diagnostic and Statistics Manual (DSM IV) criteria for ADHD. The inattention, hyperactivity and impulsiveness items formed distinct constructs and were analysed separately. Teachers were asked to rate Grade 1 children on a six-point scale (0 to 5) covering the range of never to always. In our example, we look at the Rasch statistics obtained for the nine items which form the construct *inattention*.

Global model and item fit statistics

Examine the inattention construct functioning as a whole in Table 2 by examining the model fit, reliability coefficients and indices, as well as the general item, mean and person fit.

Table 2. General instrument and item functioning: model fit, reliability, mean item and person fit

Winsteps	**RUMM**
Overall Model Fit Log-likelihood chi-squared: 54693.34, probability = .985 **Interpretation Model Fit:** a non-significant *p*-value (p>0.05) indicates the data fit the Rasch model **Reliability statistics** Real Person: 0.94, Separation: 3.84 Real Item: 0.93, Separation: 3.73 **Interpretation Reliability:** The person reliability is excellent, and the separation indicates items are sensitive enough to distinguish different groups. Item reliability is excellent, and the sample is large and diverse enough to confirm item hierarchy. **Overall Person and Item Fit** Real Person RMSE: 0.59 Real Item RMSE: 0.03 **Interpretation Overall fit:** person fit = reasonable and item fit = good	**Overall Model Fit** χ^2 = 556.134 p = 0.000 **Interpretation Model Fit:** a significant result (p<0.05) indicates the data did not fit the Rasch model **Reliability statistics** Person Separation Index: 0.948 Cronbach Alpha: 0.9679 **Power of Analysis of Fit:** Excellent **Interpretation Reliability:** Both the PSI and the Cronbach's alpha are excellent and indicate that the items can distinguish between a range of abilities and persons. **Overall Person and Item Fit** Person Fit Residual: -0.795, *SD* = 1.889 Item Fit Residual: -0.340, *SD* = 9.203 **Interpretation Overall fit:** item fit residual in the ideal range, but the *SD* is larger than desirable

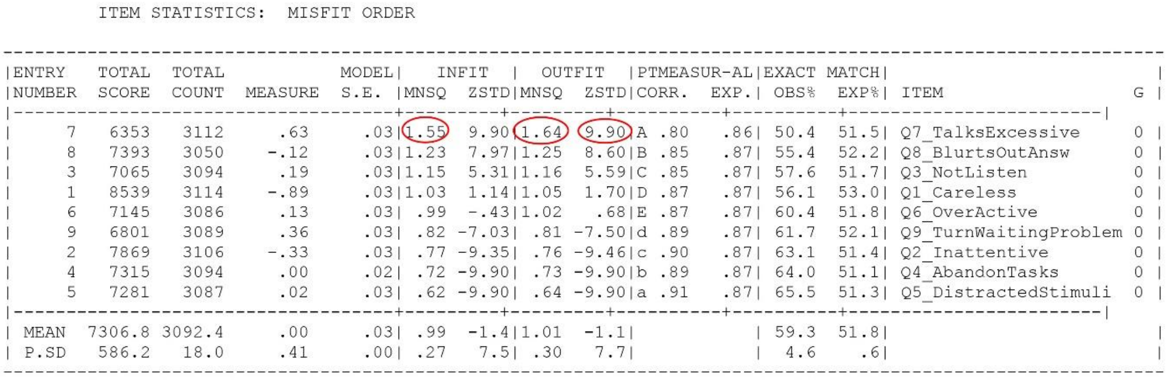
Winsteps shows a non-significant chi-square, indicating that there is no significant difference between the model and the data. But RUMM does show a significant chi-square. The two programmes use different methods to fit the data to the Rasch model[12]. Therefore, some disparities between the two types of outputs are expected. General model fit is less important

[12] Winsteps uses joint maximum likelihood estimation (JMLE); newer versions offer Conditional maximum likelihood estimation (CMLE). RUMM utilises pairwise conditional maximum likelihood

than individual item fit. Consequently, we take note of the significant χ^2 reported by RUMM, but we do not draw any conclusions until we have examined the other outputs. Both Winsteps and RUMM show excellent reliability indices for persons and items, a good indication of sufficient sample size and as well as item hierarchy and range. Residual estimates indicate data to model fit, whether the expected person ability location or item difficulty is the same as the observed scores. The general item and person fit indicated by the RMSE in Winsteps shows that person fit is reasonable, and item fit is good. The RUMM fit statistics indicate that item fit residual is within the expected range, but the *SD* is larger than preferable. The fit residual for the persons is just outside the expected range, and the person *SD* is very large, indicating a wide range of abilities.

Items misfit and category and threshold functioning

Winsteps and RUMM both provide item residuals, which are similar and any differences observed are due to the estimation procedures (Tennent & Pallant, 2006). The standardised residuals are equivalent to effect sizes; recommendations such as values above/below 1.5 or 2.5 indicate more than a standard deviation away from the mean. Such recommendations are rules of thumb, and the user should consider whether their assessment is a high stake and which criteria best fits the aims of their assessment. Winsteps indicates that one item has an MNSQ infit and outfit above 1.5, but seven items have large ZSTD values above or below 2, which is shown in Figure 8. Based on the Winsteps results, it would be prudent to examine item 7 with the larger than desirable MNSQ. Outfit and infit statistics are sensitive to unexpected responses when items are easier or more difficult than expected by the model. Both may indicate underfit and variance not accounted for by the model and should be investigated.

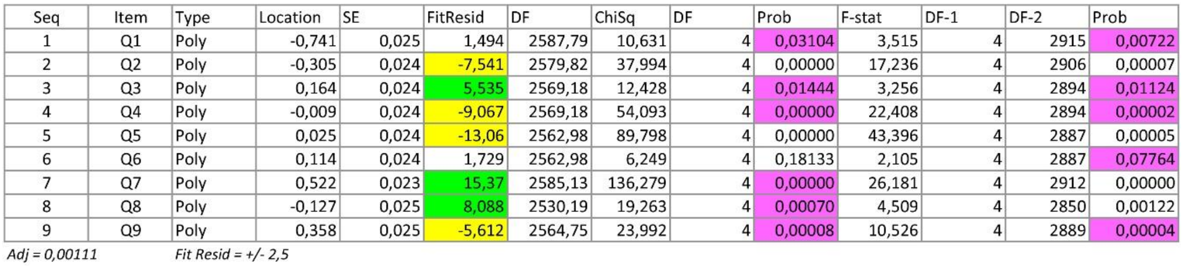


```
        ITEM STATISTICS:  MISFIT ORDER

-------------------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PTMEASUR-AL|EXACT MATCH|                    |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM             G |
|------------------------------------+----------+----------+-----------+-----------+-------------------|
|     7   6353   3112     .63    .03|1.55  9.90|1.64  9.90|A .80   .86| 50.4  51.5| Q7_TalksExcessive     0 |
|     8   7393   3050    -.12    .03|1.23  7.97|1.25  8.60|B .85   .87| 55.4  52.2| Q8_BlurtsOutAnsw      0 |
|     3   7065   3094     .19    .03|1.15  5.31|1.16  5.59|C .85   .87| 57.6  51.7| Q3_NotListen          0 |
|     1   8539   3114    -.89    .03|1.03  1.14|1.05  1.70|D .87   .87| 56.1  53.0| Q1_Careless           0 |
|     6   7145   3086     .13    .03| .99  -.43|1.02   .68|E .87   .87| 60.4  51.8| Q6_OverActive         0 |
|     9   6801   3089     .36    .03| .82 -7.03| .81 -7.50|d .89   .87| 61.7  52.1| Q9_TurnWaitingProblem 0 |
|     2   7869   3106    -.33    .03| .77 -9.35| .76 -9.46|c .90   .87| 63.1  51.4| Q2_Inattentive        0 |
|     4   7315   3094     .00    .02| .72 -9.90| .73 -9.90|b .89   .87| 64.0  51.1| Q4_AbandonTasks       0 |
|     5   7281   3087     .02    .03| .62 -9.90| .64 -9.90|a .91   .87| 65.5  51.3| Q5_DistractedStimuli  0 |
|------------------------------------+----------+----------+-----------+-----------+-------------------|
| MEAN  7306.8 3092.4     .00    .03| .99  -1.4|1.01  -1.1|           | 59.3  51.8|                    |
| P.SD   586.2   18.0     .41    .00| .27   7.5| .30   7.7|           |  4.6    .6|                    |
-------------------------------------------------------------------------------------------------------
```

Figure 8. Item fit order in Winsteps (Source: Author)

RUMM indicates that 5 out of the 9 items have large fit residuals (highlighted in yellow/green) which are also significant, as shown below in Figure 9. All five questions should be examined for the category and threshold functioning, DIF, discrimination and spread.



Seq	Item	Type	Location	SE	FitResid	DF	ChiSq	DF	Prob	F-stat	DF-1	DF-2	Prob
1	Q1	Poly	-0,741	0,025	1,494	2587,79	10,631	4	0,03104	3,515	4	2915	0,00722
2	Q2	Poly	-0,305	0,024	-7,541	2579,82	37,994	4	0,00000	17,236	4	2906	0,00007
3	Q3	Poly	0,164	0,024	5,535	2569,18	12,428	4	0,01444	3,256	4	2894	0,01124
4	Q4	Poly	-0,009	0,024	-9,067	2569,18	54,093	4	0,00000	22,408	4	2894	0,00002
5	Q5	Poly	0,025	0,024	-13,06	2562,98	89,798	4	0,00000	43,396	4	2887	0,00005
6	Q6	Poly	0,114	0,024	1,729	2562,98	6,249	4	0,18133	2,105	4	2887	0,07764
7	Q7	Poly	0,522	0,023	15,37	2585,13	136,279	4	0,00000	26,181	4	2912	0,00000
8	Q8	Poly	-0,127	0,025	8,088	2530,19	19,263	4	0,00070	4,509	4	2850	0,00122
9	Q9	Poly	0,358	0,025	-5,612	2564,75	23,992	4	0,00008	10,526	4	2889	0,00004

Adj = 0,00111 *Fit Resid = +/- 2,5*

Figure 9. RUMM individual item fit statistics (Source: Author)

Take note that in both Winsteps and RUMM, item 7 (Loses Equipment) is the most misfitted item and significantly so in both software packages. The five items in RUMM with large fit residuals are also items with large standard scores in Winsteps. In Winsteps, the outfit MNSQ would be considered first, then the Infit MSNQ and only after that the ZSTD. The infit (inliers) and outfit (outliers) mean-squares (MNSQ) are the squared standardised fit residuals. Interpreting the ZSTD with large sample sizes is not recommended as it is sensitive to sample size and stringent on exact model fit. In RUMM, items with both large fit residuals (less than

-2.5 or more than +2.5), which are also significant (Bonferroni adjusted *p*-value) are considered potentially misfitted.

In terms of categories, category probability curves in both Winsteps and RUMM indicate that categories 0 (never) and 5 (always) highly endorsed for some items and hardly at all for other items. The example of item 9 (forgetful) shown in Figure 10**Error! Reference source not found.** and Figure 11 indicates that category 5 is very close to category 4, and categories in the middle are less likely to be endorsed.

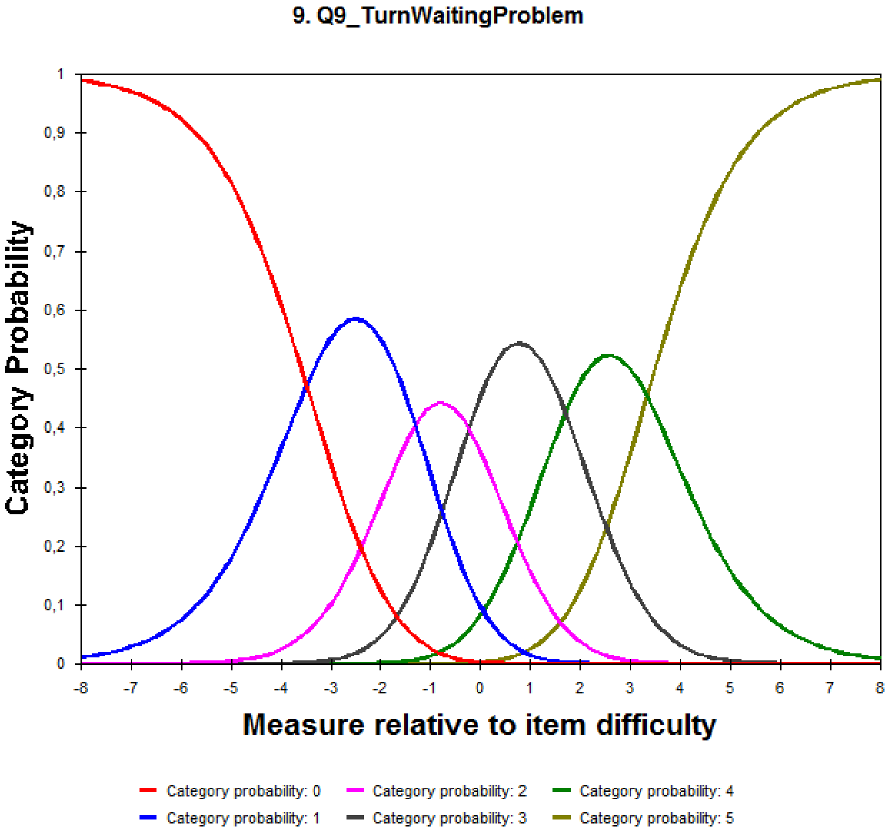


Figure 10. Item 9 category curves Winsteps (Source: Author)

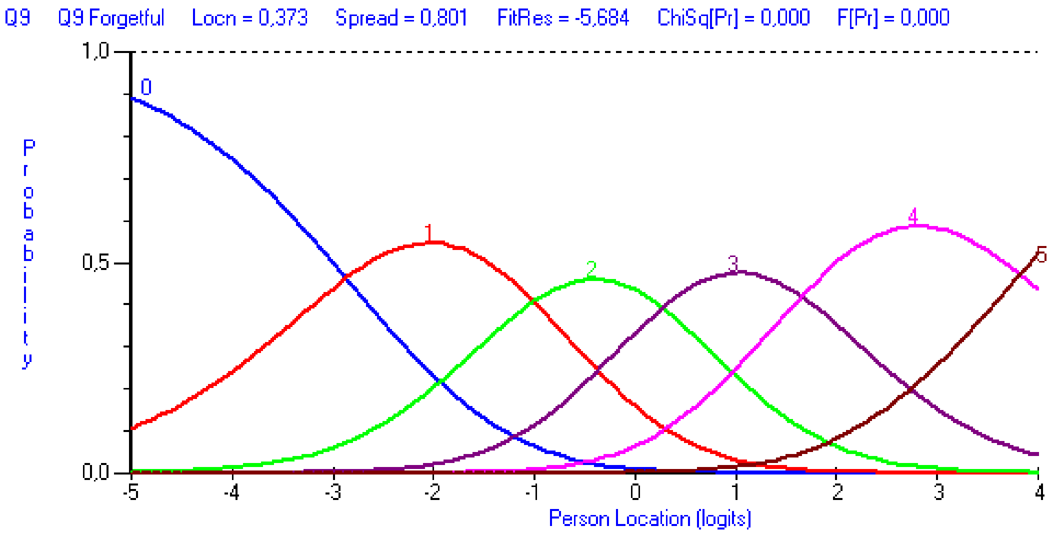
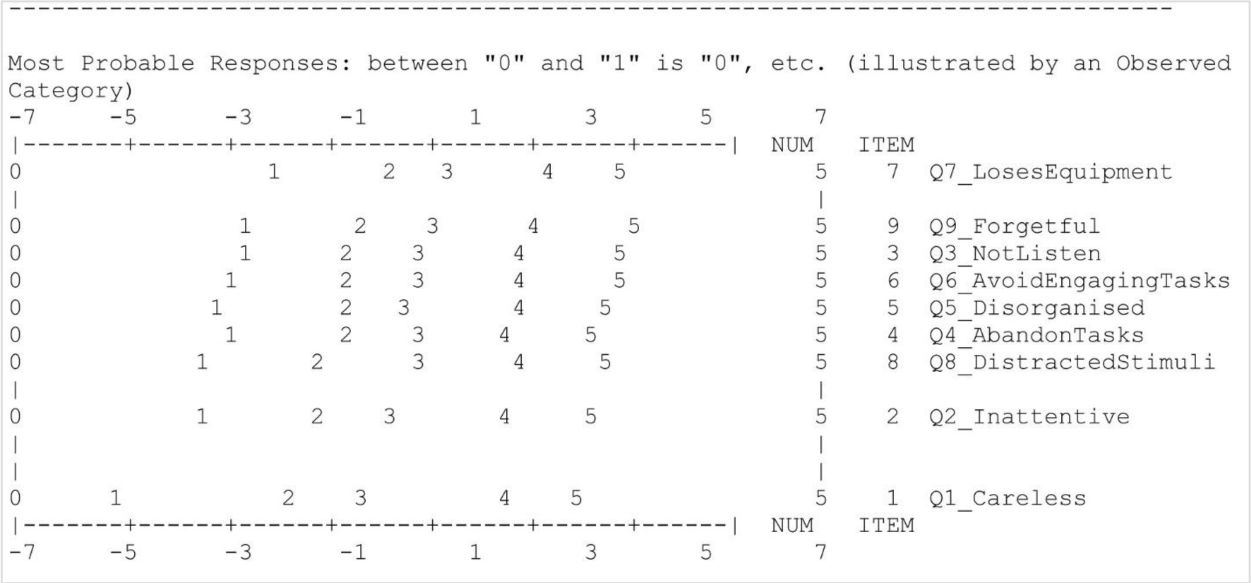


Figure 11. Item 9 category curves RUMM (Source: Author)

Item fit could be improved if some of the middle categories are collapsed (combined). However, carefully try this out as it will reduce the person reliability index. Consider the implications for the construct and measurement before collapsing categories. An ideal item would have categories that increase equally and orderly so that each category contributes to measurement. In Winsteps, investigate the Item-structure ISFILE to see if measurement (logit-value) increases with each subsequent category and if there is large MNSQ infit or outfit per category. In RUMM, examine the chi-square which should be non-significant and look at category coordinates. Consider category functioning in conjunction with threshold information. In the current example, none of the categories were problematic enough to warrant collapsing them.

The threshold maps of Winsteps (see Figure 12) and RUMM (Figure 13) are shown below. There are no disordered thresholds, and the probability of a category being endorsed does increase with a higher rating on the inattention construct. Each of the numbers (1, 2, 3, 4 and 5) represent a category, and the logit scale of person location is shown above and below in the graph, here as -7 to +5, see also Figure 7 for more information on how to interpret probabilities of categories.



```
---------------------------------------------------------------------------------

Most Probable Responses: between "0" and "1" is "0", etc. (illustrated by an Observed
Category)
-7     -5      -3      -1        1       3       5       7
|-------+------+------+------+------+------+------|  NUM   ITEM
0                 1       2   3      4    5             5    7  Q7_LosesEquipment
|                                                       |
0               1       2    3      4      5            5    9  Q9_Forgetful
0               1      2    3      4      5             5    3  Q3_NotListen
0              1       2    3      4      5             5    6  Q6_AvoidEngagingTasks
0             1        2   3       4     5              5    5  Q5_Disorganised
0              1       2    3     4     5               5    4  Q4_AbandonTasks
0            1       2      3      4     5              5    8  Q8_DistractedStimuli
|                                                       |
0            1       2    3       4     5               5    2  Q2_Inattentive
|                                                       |
|                                                       |
0      1           2    3         4    5                5    1  Q1_Careless
|-------+------+------+------+------+------+------|  NUM   ITEM
-7     -5      -3      -1        1       3       5       7
```

Figure 12. Threshold map Winsteps (Source: Author)

However, on the threshold maps, we can see that not all categories are equally spread, as shown in the RUMM threshold map in Figure 13. Look at question 9; category 5 barely registers here, and this is also evident on the Winsteps map. Thresholds are not as important as categories, and the category probability curves should be examined in conjunction with the threshold maps.

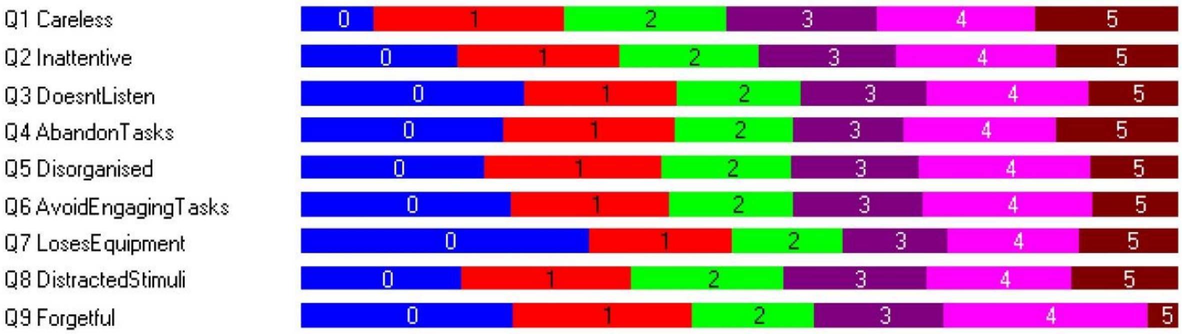
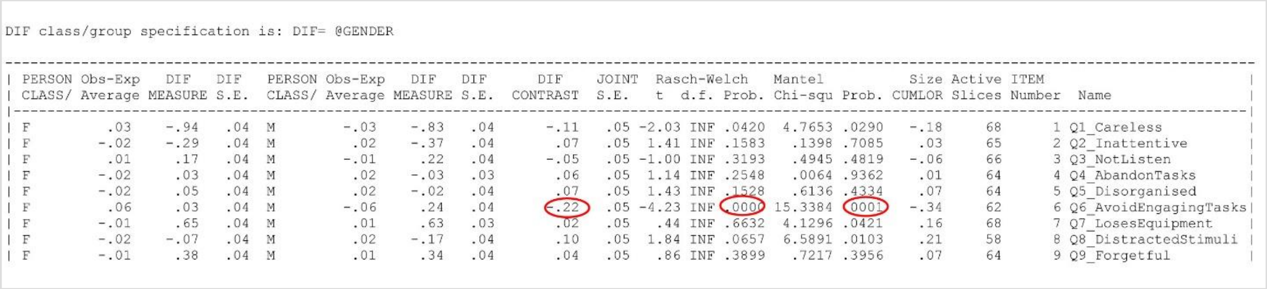


Figure 13. Threshold map RUMM (Source: Author)

Invariance via differential item functioning

Measurement invariance is a multifaceted aspect of instruments that can be examined using a variety of statistics. But for this introduction, let us look at the most often used gauge, differential item functioning (DIF). In Winsteps, we look at Table 30 to see if boys or girls were more likely to receive higher or lower inattention ratings. DIF is present if they were scored differently despite having the same underlying rating on the construct (see Figure 14).



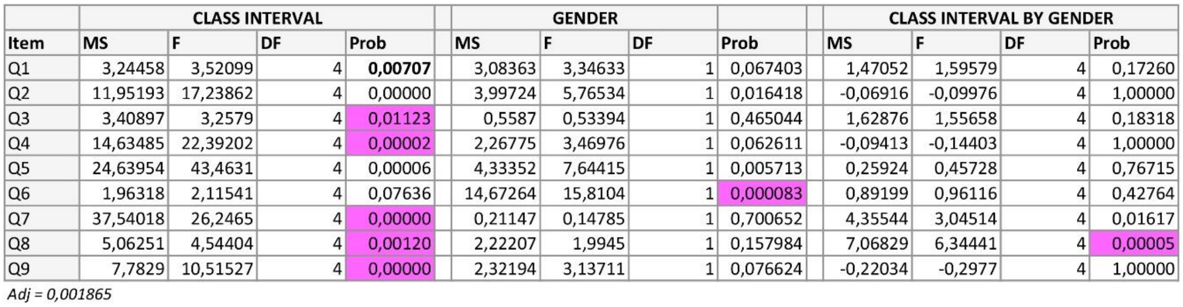
```
DIF class/group specification is: DIF= @GENDER

-------------------------------------------------------------------------------------------------------------------------------------------------
| PERSON Obs-Exp    DIF   DIF    PERSON Obs-Exp    DIF   DIF       DIF    JOINT  Rasch-Welch    Mantel          Size Active ITEM                        |
| CLASS/ Average MEASURE S.E.   CLASS/ Average MEASURE S.E.  CONTRAST  S.E.    t  d.f. Prob. Chi-squ Prob.  CUMLOR Slices Number  Name                   |
|------------------------------------------------------------------------------------------------------------------------------------------------|
| F          .03    -.94   .04  M          -.03    -.83   .04       -.11   .05 -2.03 INF .0420  4.7653 .0290   -.18     68        1 Q1_Careless            |
| F         -.02    -.29   .04  M           .02    -.37   .04        .07   .05  1.41 INF .1583   .1398 .7085    .03     65        2 Q2_Inattentive         |
| F          .01     .17   .04  M          -.01     .22   .04       -.05   .05 -1.00 INF .3193   .4945 .4819   -.06     66        3 Q3_NotListen           |
| F         -.02     .03   .04  M           .02    -.03   .03        .06   .05  1.14 INF .2548   .0064 .9362    .01     64        4 Q4_AbandonTasks        |
| F         -.02     .05   .04  M           .02    -.02   .04        .07   .05  1.43 INF .1528   .6136 .4334    .07     64        5 Q5_Disorganised        |
| F          .06     .03   .04  M          -.06     .24   .04       -.22   .05 -4.23 INF .0000 15.3384  0001   -.34     62        6 Q6_AvoidEngagingTasks|
| F         -.01     .65   .04  M           .01     .63   .03        .02   .05   .44 INF .6632  4.1296 .0421    .16     68        7 Q7_LosesEquipment      |
| F         -.02    -.07   .04  M           .02    -.17   .04        .10   .05  1.84 INF .0657  6.5891 .0103    .21     58        8 Q8_DistractedStimuli  |
| F         -.01     .38   .04  M           .01     .34   .04        .04   .05   .86 INF .3899   .7217 .3956    .07     64        9 Q9_Forgetful           |
```

Figure 14. Differential Item Functioning Winsteps

As can be seen in Figure 14, boys have a significantly (*p*<0.05) higher rating than girls on item 6 (avoids engaging tasks). However, the contrast of .22 is smaller than .50, and the effect is not large enough to warrant investigation in a Winsteps analysis. Figure 14 shows the uniform DIF. To obtain results for non-uniform DIF, examine the empirical Item Characteristic Curve (ICC) in the graphs window for an initial indication of non-uniform DIF. In the dialogue box, when selecting Table 30, add the specification of MA and number of ability levels to compare groups. In our example use: DIF = @GENDER+MA3 to compare three ability levels. Then look at Table 30.2 and the plots to identify non-uniform DIF.

In Figure 15, the RUMM DIF results, which indicate significant differences between boys and girls rating on the inattention scale, are highlighted in pink after the Bonferroni correction.



	CLASS INTERVAL				GENDER				CLASS INTERVAL BY GENDER			
Item	MS	F	DF	Prob	MS	F	DF	Prob	MS	F	DF	Prob
Q1	3,24458	3,52099	4	**0,00707**	3,08363	3,34633	1	0,067403	1,47052	1,59579	4	0,17260
Q2	11,95193	17,23862	4	0,00000	3,99724	5,76534	1	0,016418	-0,06916	-0,09976	4	1,00000
Q3	3,40897	3,2579	4	0,01123	0,5587	0,53394	1	0,465044	1,62876	1,55658	4	0,18318
Q4	14,63485	22,39202	4	0,00002	2,26775	3,46976	1	0,062611	-0,09413	-0,14403	4	1,00000
Q5	24,63954	43,4631	4	0,00006	4,33352	7,64415	1	0,005713	0,25924	0,45728	4	0,76715
Q6	1,96318	2,11541	4	0,07636	14,67264	15,8104	1	0,000083	0,89199	0,96116	4	0,42764
Q7	37,54018	26,2465	4	0,00000	0,21147	0,14785	1	0,700652	4,35544	3,04514	4	0,01617
Q8	5,06251	4,54404	4	0,00120	2,22207	1,9945	1	0,157984	7,06829	6,34441	4	0,00005
Q9	7,7829	10,51527	4	0,00000	2,32194	3,13711	1	0,076624	-0,22034	-0,2977	4	1,00000

Adj = 0,001865

Figure 15. Differential Item Functioning RUMM (Source: Author)

The RUMM results shows significant uniform DIF for gender, item 6 (avoids engaging tasks), while there is evidence for non-uniform DIF for item 8 (distracted by external stimuli). Further investigation of the items is needed to understand the origin of the DIF. Significant DIF could indicate a potential bias for the item, or alternatively, the construct of inattention functions differently for boys and for girls. If the second reason holds true, it may be that the construct should be analysed separately for girls and boys, and they would then be viewed as non-comparable samples.

The DIF graphs from Winsteps are shown in Figure 16. Winsteps also provides Mantel Chi-square probabilities to check the significance of the differences.

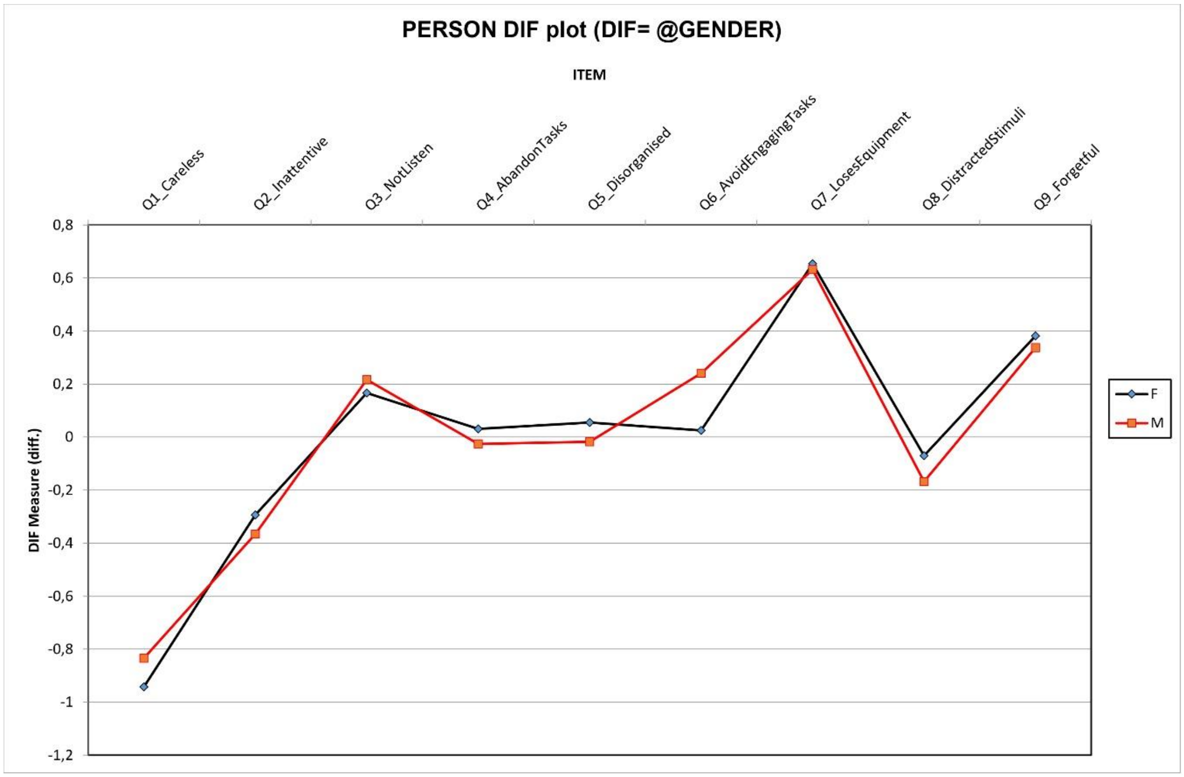


Figure 16. Item 6 DIF Winsteps (Source: Author)

Figure 16 shows differences for items 6 and 8. How much DIF is too much? In a well-functioning instrument, we prefer no items to display large and significant DIF if the construct functions in the same way for different groups. However, some DIF could be tolerated if it does not threaten the overall invariance of the construct (Combrinck, 2020).

Unidimensionality and local independence of items

In Table 3, criteria for unidimensionality is shown as well as the interpretation of the statistics produced by Winsteps and RUMM.

Table 3. Dimensionality statistics from Winsteps and RUMM

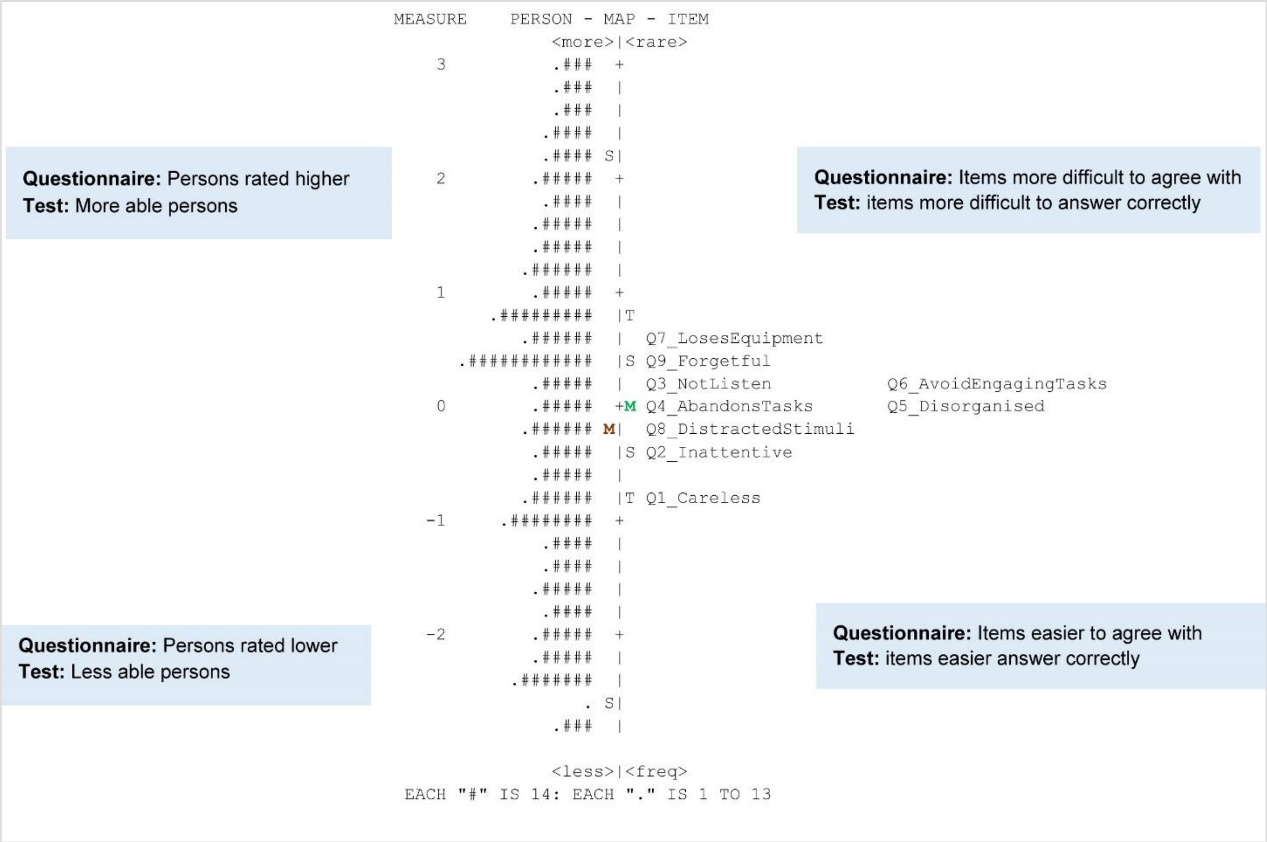
Winsteps	RUMM
PCA eigenvalues	**PCA eigen values** Largest PCA = 1.833

Unexplained variance in 1st contrast = 1.7820 **Interpretation PCA:** Acceptable, no secondary dimensions, eigenvalues lower than 2.0 **Local independence of items – correlations** Correlations ranged between .16 to .30, according to Winsteps criteria this is well below .70. There is evidence of local independence of items, and the invariance of the instrument is not threatened.	**Interpretation PCA:** Acceptable, no secondary dimensions, eigenvalues lower than 2.0 **Local independence of items – correlation matrix** Correlations ranged between .029 to .313, applying RUMM criteria there is one correlation above .300: Item 7 (Loses Equipment) and Q2 (Inattentive) have a correlation of .313. Do these two items measure the same thing and provide redundant information? Or form a construct on their own? Examining the construct and consulting subject matter experts, the conclusion is that the correlation may be spurious, and both items are needed to measure the construct.

The conclusion from both the Winsteps and RUMM analysis is that the assumption of unidimensionality for the inattention construct holds. No secondary constructs are being measured.

Person and item alignment

In theory, we want the construct to be measured comprehensively, without construct over or under-representation. Additionally, we want the range of person abilities or agreement on the construct to be measured in a comprehensive range. The Wright-map gives us a good indication of the item hierarchy and how this aligns with the spread of people on the construct (see Figure 17 below).



```
MEASURE     PERSON - MAP - ITEM
                 <more>|<rare>
    3            .###  +
                 .###  |
                 .###  |
                .####  |
                .#### S|
    2          .#####  +
                .####  |
               .#####  |
               .#####  |
              .######  |
    1          .#####  +
           .#########  |T
              .######  |   Q7_LosesEquipment
        .############  |S Q9_Forgetful
               .#####  |   Q3_NotListen           Q6_AvoidEngagingTasks
    0          .#####  +M Q4_AbandonsTasks       Q5_Disorganised
              .###### M|   Q8_DistractedStimuli
               .#####  |S Q2_Inattentive
               .#####  |
              .######  |T Q1_Careless
   -1       .########  +
                .####  |
                .####  |
               .#####  |
                .####  |
   -2          .#####  +
               .#####  |
             .#######  |
                    . S|
                 .###  |

                 <less>|<freq>
 EACH "#" IS 14: EACH "." IS 1 TO 13
```

Questionnaire: Persons rated higher
Test: More able persons

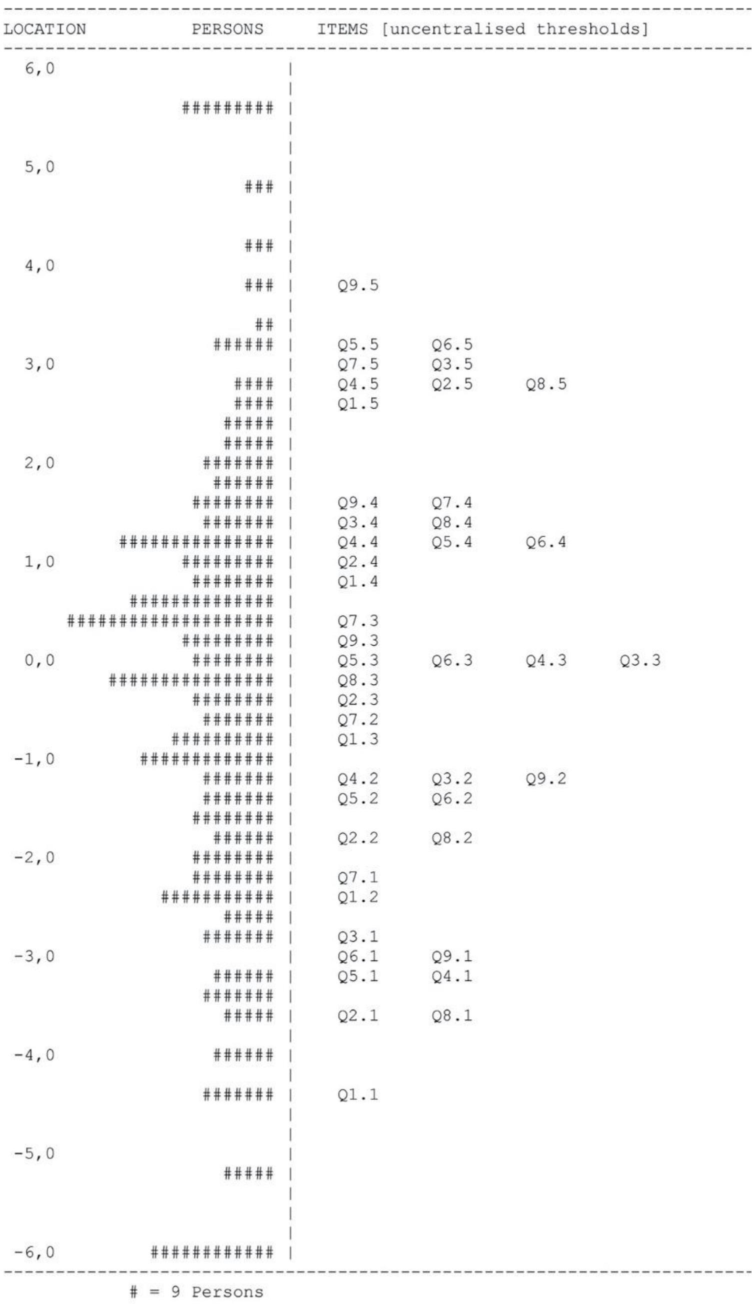
Questionnaire: Items more difficult to agree with
Test: items more difficult to answer correctly

Questionnaire: Persons rated lower
Test: Less able persons

Questionnaire: Items easier to agree with
Test: items easier answer correctly

Figure 17. Wright item-map from Winsteps (Source: Author)

On the item-map, we see that the mean of the items (green) is slightly above the mean of the persons rated on the inattention score (orange). Generally, we would want item and person means to be close. If the mean of the items is higher than the mean of the persons, the test or questionnaire may have been too difficult. Look also at the ordering of the items; question 7 (loses equipment) was the most challenging item for teachers to rate highly. Conversely, question 1 (careless) is the easiest item for teachers to agree with when rating the children. To evaluate this order, you need to understand your construct well. In Figure 18, the threshold item-map is shown, where the thresholds between 1 to 5 are located in comparison with the persons.



```
------------------------------------------------------------------------
LOCATION          PERSONS     ITEMS [uncentralised thresholds]
------------------------------------------------------------------------
  6,0                      |
                           |
                 ######### |
                           |
                           |
  5,0                      |
                       ### |
                           |
                           |
                       ### |
  4,0                      |
                       ### |    Q9.5
                           |
                        ## |
                    ###### |    Q5.5     Q6.5
  3,0                      |    Q7.5     Q3.5
                      #### |    Q4.5     Q2.5     Q8.5
                      #### |    Q1.5
                     ##### |
                     ##### |
  2,0              ####### |
                    ###### |
                  ######## |    Q9.4     Q7.4
                   ####### |    Q3.4     Q8.4
           ############### |    Q4.4     Q5.4     Q6.4
  1,0            ######### |    Q2.4
                  ######## |    Q1.4
            ############## |
      #################### |    Q7.3
                 ######### |    Q9.3
  0,0             ######## |    Q5.3     Q6.3     Q4.3     Q3.3
          ################ |    Q8.3
                  ######## |    Q2.3
                   ####### |    Q7.2
                ########## |    Q1.3
 -1,0        ############# |
                   ####### |    Q4.2     Q3.2     Q9.2
                   ####### |    Q5.2     Q6.2
                  ######## |
                    ###### |    Q2.2     Q8.2
 -2,0             ######## |
                  ######## |    Q7.1
               ########### |    Q1.2
                     ##### |
                   ####### |    Q3.1
 -3,0                      |    Q6.1     Q9.1
                    ###### |    Q5.1     Q4.1
                   ####### |
                     ##### |    Q2.1     Q8.1
                           |
 -4,0               ###### |
                           |
                   ####### |    Q1.1
                           |
                           |
 -5,0                      |
                     ##### |
                           |
                           |
                           |
 -6,0         ############ |
------------------------------------------------------------------------
            # = 9 Persons
```

Figure 18. Threshold item-map from RUMM (Source: Author)

The ordering in the threshold map reflects what we would expect to see, ratings of 5 are the most difficult to agree with, and ratings of 1 are the easiest. If the map did not reflect this neat ordering, there would be a problem with the thresholds and item ordering.

CONCLUSION

Measurement is a science and instruments need to adhere to the principles of measurement if an assessment is reliable and valid. Social science instruments, such as tests, questionnaires and checklists, should be evaluated with psychometric models to judge if they sufficiently provide scientific inferences. Rasch theory offers various one-parameter models which convert raw item scores to an interval scale. The logit scale can be extracted from the software, but if the items function well enough with regards to the criteria discussed, the raw scores will be highly correlated to the logit scale. In essence, the Rasch models test the hypothesis that accurate and useful measurement is evident in an instrument, and if the hypothesis is supported, the raw scores are sufficient.

The models generate statistics which can be used to gauge measurement invariance, unidimensionality, item functioning and ordering representative of the underlying construct. The current chapter provided an introduction to the use and interpretation of the Rasch measurement model outputs for evaluating an existing instrument or designing a new instrument. This chapter discussed the scientific and ethical need for decolonising measurement and the ways in which Rasch theory can be used to achieve this goal. It emphasised the need for the Africanisation of instruments. Applying psychometric theories, such as Rasch and IRT, is offered as one of many avenues for improving assessment and increasing ethical standards in the social sciences.

The chapter assumes basic knowledge of statistics and levels of measurement to understand the application of the Rasch measurement model. The Rasch statistics provide quantitative evidence of the internal functioning of a test or questionnaire. The reader is encouraged to also apply qualitative methods to the design of a new instrument. The Rasch dichotomous and polytomous models were briefly discussed. The emphasis was on the understanding that there are different models to accommodate all types of items. Instruments can contain a mixture of item types and the software, Winsteps and RUMM, can apply the correct model to each item. The assumptions of Rasch models align with the principles of measurement. Measurement principles include equal intervals, additivity and concatenation, unidimensionality, invariance and order. When items fail to meet the assumptions, they should be refined (rephrased/retranslated/restructured) or possibly excluded if they cannot be improved.

When conducting your Rasch analysis, it is recommended that you structure the analysis so that you properly evaluate the internal functioning of your instrument. Evaluating your test or questionnaire should include reliability indices, fit statistics and unidimensionality. You should also investigate invariance, category and threshold functioning and degree of alignment of persons and items. The chapter presented steps to prepare your data for Winsteps and RUMM, as well as interpretation guidelines for the outputs. The emphasis is on interpreting all the various outputs holistically. Decisions about items should be made based on the operationalisation of the construct. The theory of the construct should guide any changes made to the instrument; the statistical output is secondary to the qualitative understanding of the underlying trait.

The current chapter offers a brief introduction to Rasch measurement theory, and there are many topics not covered, such as the item and test information functioning, anchoring and equating. Resources for further topics and more details are shown below, with the online courses recommended for the user who wants more practical inputs. Examples of Rasch applications in South Africa has been added in Figure 19 to demonstrate the usefulness of the model for our context. Kagee and De Bruin (2007) started with qualitative interviews to understand the construct of *detainee distress* and developed their items based on themes. Thereafter they piloted the instrument and used Rasch models to refine the instrument. Makhubela (2019) used Rasch models and confirmatory factor analysis (CFA) to examine the applicability of the *Trauma Symptom Checklist for Children* in the South African context. Schutte et al. (2019) examined the *Satisfaction with Life Scale* in South Africa by applying Rasch models and comparing results with Italian responses to make recommendations for improving the instrument. The novice user of Rasch models is encouraged to utilise the resources presented in Figure 19.

Introductory Books

Applying the Rasch Model: Fundamental Measurement in the Human Sciences (Ed. 3), by Bond and Fox (2015)

Rasch analysis in the human sciences by Boone, Staver & Yale (2014)

A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences by Andrich & Marais (2019)

Online Resources

Research Papers, Explorations & Explanations: https://www.rasch.org/rasch.htm

Rasch Measurement Transactions: https://www.rasch.org/rmt/index.htm

John Michael Linacre YouTube channel: https://www.youtube.com/channel/UCVgROegi_QOBkZ7uJeAY4sg

Discussion Forums

Rasch Measurement Forum: https://raschforum.boards.net/

The Matilda Bay Club: join email group at rasch@wu.ac.at

Online Courses

Practical Rasch Measurement Workshop: https://www.statistics.com/courses/rasch-measurement-core-topics/

An Introduction to Rasch Measurement Theory and RUMM2030 Course: http://www.education.uwa.edu.au/ppl/courses

Examples of South African applications of Rach Models

Kagee, A., & De Bruin, G. P. (2007). The South African former detainees distress scale: results of a Rasch item response theory analysis. *South African Journal of Psychology, 37*(3), 518–529. https://doi.org/10.1177/008124630703700309.

Schutte, L., Wissing, M. P., Negri, L., & Delle, F. A. (2019). Rasch analysis of the satisfaction with life scale across countries: findings from South Africa and Italy. Current Psychology (2019). https://doi.org/10.1007/s12144-019-00424-5.

Makhubela, M. (2019). Using the trauma symptom checklist for children-short form (TSCC-SF) on abused children in South Africa: confirmatory factor analysis and Rasch models. *Child Abuse & Neglect, 98*.

Figure 19: Introductory Texts, Online Resources and Discussion Forums

REFERENCES

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care, 42*(1), 16.

Andrich, D., Lyne, A., Sheridan, B. & Luo, G. (2010). *RUMM 2030 [Computer Software]*. RUMM Laboratory.

Andrich. D. & Marais. I. (2019). *A Course in Rasch Measurement Theory: Measuring in the Educational. Social and Health Sciences*. Springer.

Baghaei, P. (2008). *The Rasch Model as a Construct Validation Tool. Rasch Measurement Transactions, 22*(1), 1145-1146. Downloaded from: https://www.rasch.org/rmt/rmt221a.htm.

Barnes, B., Siswana, A., Ratele, K., Cornell, J., Dlamini, S., Helman, R., & Titi, N. (2018). Some basic questions about decolonizing Africa(n)-centred psychology considered. *South African Journal of Psychology, 48*(3), 331-342. doi:10.1177/0081246318790444.

Bedell, B., van Eeden, R., & van Staden, F. (1999). Culture as a moderator variable in psychological test performance: Issues and trends in South Africa. *Journal of Industrial Psychology, 25*, 1–7.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez HR, & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in Public Health, 6,* 149–149. https://doi.org/10.3389/fpubh.2018.00149.

Bond. T. G. & Fox. C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (Ed. 3). New York: Routledge.

Boone. W. & Rogan. J. (2005). Rigour in quantitative analysis: The promise of Rasch analysis techniques. *African Journal of Research in Mathematics. Science and Technology Education. 9*(1). 25-38.

Boone. W. J. & Noltemeyer. A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Educational Psychology and Counselling, 4.* 1-13.

Boone. W. J. (2016). Rasch Analysis for Instrument Development: Why When and How? *CBE Life Science Education. 15*(4), 1-7.

Boone. W. J. Staver. J. R. & Yale. M. S. (2014). *Rasch analysis in the human sciences*. Springer.

Bradley, M., & Brand, A. (2016). Significance testing needs a taxonomy: Or how the fisher, neyman-pearson controversy resulted in the inferential tail wagging the measurement dog. *Psychological Reports, 119*(2), 487-504. doi:10.1177/0033294116662659.

Bruschi, A. (2017). Measurement in social research: Some misunderstandings. *Quality & Quantity: International Journal of Methodology, 51*(5), 2219-2243. doi:10.1007/s11135-016-0383-5.

Burton-Jones, A., & Lee, A. S. (2017). Thinking about measures and measurement in positivist research: a proposal for refocusing on fundamentals. *Information Systems Research, 28*(3), 451–467. https://doi.org/10.1287/isre.2017.0704.

Chen, W., Lenderking, W., Jin, Y., Wyrwich, K., Gelhorn, H., & Revicki, D. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? an example using PROMIS pain behavior item bank data. *Quality of Life Research, 23*(2), 485-493.

Claasen, N. C. W. (1997). Cultural differences, politics and test bias in South Africa. *European Review of Applied Psychology, 47*, 297–307.

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics.* SAGE Publications Ltd doi:10.4135/9781446221556.

Cohen, R. J., & Swerdlik, M. E. (2018). *Psychological testing and assessment: an introduction to tests and measurement (Ninth).* McGraw-Hill Education.

Combrinck, C. (2018). *The use of Rasch Measurement Theory to address measurement and analysis challenges in social science research*. (Doctoral thesis). University of Pretoria.

Combrinck, C. (2020). Big changes in achievement between cohorts: a true reflection of educational improvement or is the test to blame? In Khine, M.S. (ed.), *Rasch Measurement: Applications in Quantitative Educational Research* (pp. 179-196). Springer.

Corcoran, T. (Ed.). (2014). *Psychology in education: Critical theory~practice (Innovations and controversies)*. Rotterdam: Sense. doi:10.1007/978-94-6209-566-3.

David, S. L., Hitchcock, J. H., Ragan, B., Brooks, G., & Starkey, C. (2018). Mixing interviews and Rasch modeling: demonstrating a procedure used to develop an instrument that measures trust. *Journal of Mixed Methods Research, 12*(1), 75–94.

Engelhard, G. J. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research & Perspective, 6*(3), 155–189. https://doi.org/10.1080/15366360802197792.

Fox, C. & Jones, J. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology. 45*(1), 30-45.

Franchignoni, F., Salaffi, F., Giordano, A., Ciapetti, A., Carotti, M., & Ottonello, M. (2012). Psychometric properties of self-administered lequesne algofunctional indexes in patients with hip and knee osteoarthritis: an evaluation using classical test theory and Rasch analysis. *Clinical Rheumatology, 31*(1), 113–21. https://doi.org/10.1007/s10067-011-1788-0.

Fried, E.I. & Flake, J.F. (2018). Measurement Matters. Observer Magazine, 31. Downloaded from https://www.psychologicalscience.org/observer/measurement-matters.

Gerriet, J., Valerie, M., & Jonathan, T. (2014). Classical test theory and item response theory: two understandings of one high-stakes performance exam. *Colombian Applied Linguistics Journal, 16*(2), 167–184. https://doi.org/10.14483/udistrital.jour.calj.2014.2.a03.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Harwell, M.R., Gatti, G.G. & Linacre, J.M. (2002). "Linear" Rescaling vs. Linear Measurement. *Rasch Measurement Transactions, 16*(3), 890-891. Downloaded from: https://www.rasch.org/rmt/rmt163h.htm.

Iramaneerat, C., Smith, E.V. & Smith, R.M. (2008). An Introduction to Rasch Measurement. In Osborne, J. (Ed). *Best Practices in Quantitative Methods* (pp. 50-70). Sage Publications.

Jafari, P., Bagheri, Z., Ayatollahi, S.M.T. et al. (2012). Using Rasch rating scale model to reassess the psychometric properties of the Persian version of the PedsQLTM 4.0 Generic Core Scales in school children. *Health Quality Life Outcomes, 10*(27). https://doi.org/10.1186/1477-7525-10-27.

Kagee, A., & De Bruin, G. P. (2007). The South African former detainees distress scale: results of a Rasch item response theory analysis. *South African Journal of Psychology, 37*(3), 518–529. https://doi.org/10.1177/008124630703700309.

Krägeloh, C.U., Kersten, P., Billington, D. R., Hsu, P. H.-C., Shepherd, D., Landon, J., & Feng, X. J. (2013). Validation of the WHOQOL-Brief quality of life questionnaire for general use in New Zealand: confirmatory factor analysis and Rasch analysis. *Quality of Life Research, 22*(6), 1451–1457.

Laher, S., & Cockcroft, K. (2014). Psychological assessment in post-apartheid South Africa : The way forward. *South African Journal of Psychology, 44*(3), 303-314. doi:10.1177/0081246314533634.

Laher, S., & Cockcroft, K. (Eds.). (2013). *Psychological assessment in South Africa: Research and applications*. Wits University Press.

Linacre, J. M. (2019). *Winsteps® Rasch measurement computer program User's Guide*. Beaverton, Oregon: Winsteps.com.

Linacre, J.M. (2012, June). *Winsteps Tutorial 1*. Retrieved from http://www.winsteps.com/a/winsteps-tutorial-1.pdf

Linacre, J.M. (2020). *Winsteps® (Version 4.5.0) [Computer Software]*. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2020. Available from https://www.winsteps.com/.

Makhubela, M. (2019). Using the trauma symptom checklist for children-short form (TSCC-SF) on abused children in South Africa: confirmatory factor analysis and Rasch models. *Child Abuse & Neglect*, 98.

Marais, I. & Andrich, D. (2007). *RUMM: Rasch Unidimensional Measurement Models Simulation Studies Software*. The University of Western Australian..

Massof, R. W. (2011). Understanding Rasch and item response theory models: applications to the estimation and validation of interval latent trait measures from responses to rating scale questionnaires. *Ophthalmic Epidemiology, 18*(1), 1–19. https://doi.org/10.3109/09286586.2010.545501.

Merrell, C. & Tymms, P.B. (2005). Inattention, hyperactivity and impulsiveness: Their impact on academic achievement and progress. *British Journal of Educational Psychology, 71*, 43–56.

Merrell, C., Sayal, K., Tymms, P., & Kasim, A. (2017). A longitudinal study of the association between inattention, hyperactivity and impulsivity and children's academic attainment at age 11. *Learning and Individual Differences, 53*, 156-161. doi:10.1016/j.lindif.2016.04.003.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed.). American Council on Education.

Michell, J. (2008). Is psychometrics pathological science*? Measurement: Interdisciplinary Research and Perspectives, 6*(1-2), 7-24. https://doi.org/10.1080/15366360802035489.

Milner, K., Thatcher, A., & Donald, F. (2014). Psychological assessment for redress in South African organisations: is it just? *South African Journal of Psychology, 44*(3), 333–349. https://doi.org/10.1177/0081246314535685.

Molenaar, D., & Borsboom, D. (2013). The formalization of fairness: issues in testing for measurement invariance using subtest scores. *Educational Research & Evaluation, 19*(2/3).

Mtsatse, N. & Combrinck, C. (2018). Dialects matter: The impact of dialects and code-switching on the literacy and numeracy achievement of isiXhosa Grade 1 learners in the Western Cape. *Journal of Education, 72* (19 - 36).

Nolte, S., Coon, C., Hudgens, S., & Verdam, M. G. E. (2019). Psychometric evaluation of the promis® depression item bank: an illustration of classical test theory methods. *Journal of Patient-Reported Outcomes, 3*(1), 1–10. https://doi.org/10.1186/s41687-019-0127-0.

Pearce, J. (2018). Psychometrics in action, science as practice. *Advances in Health Sciences Education: Theory and Practice, 23*(3), 653-663. doi:10.1007/s10459-017-9789-7.

Pool, R., Montgomery, C. M., Morar, N. S., Mweemba, O., Ssali, A., Gafos, M., McCormack, S. (2010). A mixed methods and triangulation model for increasing the accuracy of adherence and sexual behaviour data: the microbicides development programme. *PLOS One, 5*(7).

Power, A., Lemay, J. & Cooke, S. (2017). Justify your answer: The role of written think aloud in script concordance testing. *Teaching and Learning in Medicine, 29*(1), 59-67. doi:10.1080/10401334.2016.1217778.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Institute for Objective Measurement.

Rasch, G. (1992). *Probabilistic Models for Some Intelligence and Attainment Tests.* MESA Press.

Recabarren, D. A., Mallinckrodt, B. & Miles, J. R. (2016). Using focus groups and Rasch item response theory to improve instrument development. *The Counseling Psychologist, 44*(2), 146-194.

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

RUMM Laboratory (2015). *RUMM2030 Getting Started Manual*. Perth: RUMM Laboratory Pty Ltd.

Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. (2017). Breaking free from the limitations of classical test theory: developing and measuring information systems scales using item response theory. *Information & Management, 54*(2), 189–203. https://doi.org/10.1016/j.im.2016.06.005.

Schutte, L., Wissing, M. P., Negri, L., & Delle, F. A. (2019). Rasch analysis of the satisfaction with life scale across countries: findings from South Africa and Italy. *Current Psychology* (2019). https://doi.org/10.1007/s12144-019-00424-5.

Sehlapelo, M., & Terre Blanche, M. (1996). Psychometric testing in South Africa: Views from above and below. *Psychology in Society, 21*, 49–59.

Shaw, F. (1991). Descriptive IRT vs. Prescriptive Rasch. *Rasch Measurement Transactions, 5*(1), 131. Downloaded from: https://www.rasch.org/rmt/rmt51f.htm.

Smith, E.V., Wakely, M.B., de Kuif, R.L. & Swartz, C.W. (2003). Optimizing Rating Scales for Self-Efficacy (and Other) Research. *Educational and Psychological Measurement 63*(3), 369–91.

Stead, G.B. (2002). The Transformation of Psychology in a Post-Apartheid South Africa: An Overview. *International Journal of Group Tensions, 31*(1), 79–102. https://doi.org/10.1023/A:1014216801376.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science, 103* (2684), 677–680.

Stevens, S. S. (1959). Measurement, Psychophysics and Utility. In Churchman, C.W. & Ratoosh, P. (Eds.). *Measurement: Definitions and Theories*. John Wiley

Tennent. A. & Pallant, J. F. (2012). The Root Mean Square Error of Approximation (RMSEA) as a supplementary statistic to determine fit to the Rasch model with large sample sizes. *Rasch Measurement Transactions, 25*(4), 1348-1349.

Tennent. A. & Pallant, J. F. (2012). Unidimensionality Matters! (A Tale of Two Smiths?). *Rasch Measurement Transactions, 20*(1), 1048-51051.

Tymms, P., Howie, S., Merrell, C., Combrinck, C. & Copping, L. (2017). *The First Year at School in the Western Cape: Growth, Development and Progress*. Centre for Evaluation and Monitoring: Durham University. Downloaded from: http://www.nuffieldfoundation.org/sites/default/files/files/Tymms%2041637%20-%20SouthAfricaFinalReport%20Oct%202017.pdf.

Wright B. D. (1997). Stevens Revisited. *Rasch Measurement Transactions, 11*(1), 552-553.

Wright B.D. (1999). *Fundamental measurement for psychology*. In S.E. Embretson & S.L. Hershberger (Eds.). The new rules of measurement: What every educator and psychologist should know. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wright, B. D., & Stone, M. (1999). *Measurement Essentials*. Delaware: Wide Range.

Wright, B.D. (1992). IRT in the 1990s: Which Models Work Best? 3PL or Rasch? *Rasch Measurement Transactions, (6)*1, 196-200.

Wright, B.D. (1993).Thinking with raw scores. *Rasch Measurement Transactions, 7*(2), 299-300.

Zhou, Y. (2019). A Mixed Methods Model of Scale Development and Validation Analysis. *Measurement: Interdisciplinary Research and Perspectives, 17*(1), 38-47. https://doi org/10.1080/15366367.2018.1479088.

Appendix: Recommended inclusions in an Assessment framework

- The variable name of the item
- Item label (wording of the question or paraphrased wording)
- Item type (multiple choice, constructed response, polytomous/Likert)
- Max marks/categories per item
- Number of options
- Key if multiple choice (MC) type item
- Values used for Likert type items
- Values used for missing data
- Reverse scoring: indication of items that require recoding
- Categories or subscales in the instrument

Table 4. Example of an assessment framework

Item ID	Item Type	Item Description	Sub-scale	Rating scale	Missing Value	Sequence	Max Score	Max Categories
Q1	Likert	Q1 Careless	Inattention	0 to 5	9	1	5	6
Q2	Likert	Q2 Inattentive	Inattention	0 to 5	9	2	5	6
Q3	Likert	Q3 DoesntListen	Inattention	0 to 5	9	3	5	6
Q4	Likert	Q4 AbandonTasks	Inattention	0 to 5	9	4	5	6
Q5	Likert	Q5 Disorganised	Inattention	0 to 5	9	5	5	6
Q6	Likert	Q6 AvoidEngagingTasks	Inattention	0 to 5	9	6	5	6
Q7	Likert	Q7 LosesEquipment	Inattention	0 to 5	9	7	5	6
Q8	Likert	Q8 DistractedStimuli	Inattention	0 to 5	9	8	5	6
Q9	Likert	Q9 Forgetful	Inattention	0 to 5	9	9	5	6
Q10	Likert	Q10 Fidgets	Hyperactivity	0 to 5	9	10	5	6

Q11	Likert	Q11 LeavesSeat	Hyperactivity	0 to 5	9	11	5	6
Q12	Likert	Q12 RunsExcessive	Hyperactivity	0 to 5	9	12	5	6
Q13	Likert	Q13 Noisy	Hyperactivity	0 to 5	9	13	5	6
Q14	Likert	Q14 OverActive	Hyperactivity	0 to 5	9	14	5	6
Q15	Likert	Q15 TalksExcessive	Hyperactivity	0 to 5	9	15	5	6
Q16	Likert	Q16 BlurtsOutAnsw	Impulsivity	0 to 5	9	16	5	6
Q17	Likert	Q17 TurnWaitingProblem	Impulsivity	0 to 5	9	17	5	6
Q18	Likert	Q18 Interrupts	Impulsivity	0 to 5	9	18	5	6