JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods Inquires

SCALE CONSTRUCTION OF THE TOWNSEND PERSONALITY QUESTIONNAIRE UTILISING THE RASCH UNIDIMENSIONAL MEASUREMENT MODEL: A MEASUREMENT OF PERSONALITY

Gary Clifford TOWNSEND
Assessment & People Services, 4 Medlar Road, Randpark Ridge, GA, South Africa

E-mail: gary@apsinc.co.za **Web page:** www.apsinc.co.za

Abstract
Scales used to measure latent traits like behavioural attitudes are typically measured using classical statistical approaches. However, treating raw scores as interval scales present a fundamental problem when developing measures. To avoid these pitfalls human measurement instruments need to be constructed using Rasch analysis. The Rasch unidimensional model is currently the only method able to transform raw data into abstract equal-interval scales. The objective being for each personality dimension to have all items fit the Rasch model well, with the more endorsable items reliably preceding more difficult to endorse items in the direction of increasing levels of the underlying latent construct. Specifically, ensuring that all the items in each measured dimension manifests construct linearity and conjoint additivity. According to this view, if the data fit the model, then a scale with linearity and conjoint additivity will have been developed.

Keywords: *Measurement, Rasch analysis, personality assessment, big-five, measurement, linearity, conjoint-additivity*

1. Introduction

The Townsend Personality Questionnaire (TPQ) was developed using the Five Factor Model of personality. The Five Factor Model, often referred to as the 'Big Five' (Ewen, 1998, p.140), represents the most widely acknowledged general model of the structure of personality (Betram and Brown, 2005). It incorporates five different variables into a conceptual model for describing personality (Popkins, 1998).

The five factor theory is among the newest models developed for describing personality and has demonstrated that it is among the most practical and applicable models available in the field of personality psychology (Digman, 1990).

The Big Five are collectively a taxonomy of personality traits. In essence, a framework for understanding which traits go together. They are an empirically based phenomenon, not a theory of personality (Srivastava, 2006). It is based on language since language

itself is the structure with which we frame and understand the world around us (Lucius, 2008).

There are various well structured psychological assessments in circulation using the big-five as the basis for their construction notably, Dr Tom Buchanan's IPIP Five Factor Personality Inventory. However, despite traditional methods demonstrating both reliability and validity when measuring personality (Surgency or Extraversion (.91); Agreeableness (.88); Conscientiousness (.88); Constancy (.91), and Intellect or Imagination (.90)), there is a fundamental gap in the way all these measures are constructed.

This gap results from the application of sophisticated statistical procedures to no more than counts of observed events or levels of performance rather than focusing on constructing *measures* of human behaviour (Bond and Fox, 2007).

Fundamentally, raw item scores are unable to factor in the necessary and prerequisite requirement for measurement namely, linearity and conjoint additivity.

As a consequence, the majority of psychological and educational instruments currently in circulation perpetuate this fundamental weakness in their designs - they confuse counts with measures (Wright & Linacre, 1989).

Fisher (2002) underscores this by pointing out that 'if we can't generalize from our data, no amount of statistical hocus pocus is going to construct meaningful results.' For this reason the TPQ uses linearity and conjoint additivity as the mathematical foundation for validating it as a personality measure.

1.1. Instrument Validation with Rasch Analysis

As early as the 1930's there was a polemic regarding how one quantifies "psychological measurement".

Norman Campbell (1940) and Stanley Stevens (1946) wrestled this issue from a purely scientific and social science perspective respectively (Linacre, 2012). Campbell insisted that measurement requires a "deliberate action", "a concatenation" (like taking steps to measure out a specific length or stacking bricks one on top of the other to measure height) and believed that trying to measure personality would be tantamount to attempting to "…concatenate people's heads!" On the flip side, Stanley (1946) elected to devise a definition of measurement by simply assigning "…numbers to objects or events according to rule" (Stevens, 1946).

It was Georg Rasch (1960, 1980) who ultimately went on to developed a simple, yet revolutionary, unidimensional model that, in Cambell's words, "concatenates heads". This Rasch model forms the framework within which assessment developers can evaluate the utility of their measures (Elliott, Fox, Beltyukova, Stone, Gunderson, and Zhang, 2006) and ensure that they are applying "…a robust model for the objective measurement of latent traits…" (Hendriks, et.al., 2012).

The Rasch model is currently the only method able to "transform raw data from the human sciences into abstract equal-interval scales" (Bond & Fox, 2001). It is a logistic item response model that independently scales both items and persons along the same underlying construct (Kahler, et. al., 2004).

This characteristic is called *parameter separation* and is unique to the Rasch model (Bond & Fox, 2007). Person-free item calibration and item-free person calibration is the

condition that makes it possible to generalise measurement beyond the specific instrument being used (Wright, 1968). In essence all items should be able to be compared one with the other despite who responds to them. This ensures that the instrument is calibrated, fixed, and linear having "…uniform meaning regardless of whom we choose to measure with them." (Wright, 1968)

Rasch (1960, 1961, 1968, and 1977) designated this measurement property *specific objectivity* and regards *separability* as the basis for the specific objectivity essential for scientific inference.

He holds that for the concept of person ability (B) and item difficulty (D) to be considered meaningful, there must exist a function of the probability of a correct answer which forms an *additive* system in the parameters for persons and items (Rasch, 1960).

Its parameters $\mathbf{B_n - D_i}$ allows this relation between person ability and item difficulty parameters to be contained in one estimation equation (Wright & Stone, 1999) without the one impacting the other.

Separating Item Comparisons from Persons. Consider the Rasch Model equation:

$$\mathbf{P_{ni}\,(x_{ni} = 1 \,/\, B_n, D_i) = \exp(B_n - D_i) \,/\, [1 + \exp(B_n - D_i)]} \qquad (1.1)$$

The basic Rasch model is a dichotomous response model "…that specifies the probability, P, that person *n* of ability B_n, succeeds on item *i* of difficulty D_i (Linacre, 2012). $\mathbf{P_{ni}}$ is the probability of any person ***n*** on item ***i*** endorsing a correct (x=1) response rather than an incorrect (x=0) one, given propensity to endorse (B_n) and item endorsability (D_i). This specification is sufficient and necessary for measurement to occur (Wright & Stone, 1999).
Referencing E*quation 1.1*, one can express the odds that person ***n*** endorses item ***i*** positively as:

$$\mathbf{[P_{ni} \,/\, (1 - P_{ni})] = \exp(B_n - D_i)} \qquad (1.2)$$

In log-odds units or "logits" format *Equation 1.2 is expressed as follows:*

$$\mathbf{\log_e[P_{ni} \,/\, (1 - P_{ni})] = B_n - D_i} \qquad (1.3)$$

Leading on from the aforementioned, the equivalent log-odds for any other item ***j*** and the same person ***n*** can be expressed as:

$$\mathbf{\log_e[P_{nj} \,/\, (1 - P_{nj})] = B_n - D_j} \qquad (1.4)$$

By subtracting *Equation 1.3* from *Equation 1.4* it becomes patently clear that items ***i*** and ***j*** can now be *contained in one estimation without interference* from $\mathbf{B_n}$ or any other $\mathbf{B_m}$ producing the following:

$$\mathbf{(B_n - D_j) - (B_n - D_i) = (D_i - D_j) = \log_e\{[P_{nj}\,(1 - P_{ni})] \,/\, [P_{ni}\,(1 - P_{nj})]\}} \qquad (1.5)$$

Equation 1.5 now expresses the unique parameter separation expectation where $\mathbf{B_n}$ is completely excluded as Thurstone called for in 1928. Noticeably, $\mathbf{B_n}$ cancels out leaving the

comparison $(D_i - D_j)$ of items ***i*** and ***j*** completely unimpeded by *person* effects.

Separating Person Comparisons from Items. Referencing E*quation 1.1*, one can express the odds that person ***m*** endorses item ***i*** positively as:

$$\log_e[P_{mi} / (1 - P_{mj})] = B_m - D_i \quad (1.6)$$

In much the same way as illustrated in *Separating Item Comparisons from Persons,* person ***n*** and ***m*** can be compared by subtracting *Equation 1.6* from *Equation 1.4*:

$$(B_n - D_i) - (B_m - D_i) = (B_n - B_m) = \log_e\{[P_{ni}(1 - P_{mi})]/[P_{mi}(1 - P_{ni})]\} \quad (1.7)$$

Again, the unique parameter separation of the Rasch model enables the combination of them in *Equation 1.7* so that D_i cancels out leaving the relationship $(B_n - B_m)$ of persons ***n*** and ***m*** completely unhindered by *item* effects.

Consequently, "test-free person measurement" and "sample-free item calibration" is possible given the equations for B_n are not affected by the effects of a particular D_i and equations for D_i are unaffected by the effects of a particular B_n respectively (Wright, 1968)

Model Fit and Uni-dimensionality. As opposed to convention where "…parameters are modified and accepted or rejected based on how well they fit the data" (Bryan S.K. Kim, et al, 2004), Rasch measurement is about producing data that fit the Rasch model's specification (Bond & Fox, 2007). Within this context, the concept of fit and uni-dimensionality is inextricably bound.

Uni-dimensionality, is one of the most implicit principles underlying measurement (Bond & Fox, 2007). Wright and Linacre (1989) in fact go as far as stating that "Uni-dimensionality is an essence of measurement." Rasch measurement requires this concept of a single underlying uni-dimensional variable on the data.

Because uni-dimensionality, in practice, is an abstraction rather than quantitative it is understandable that there can be no measure that is perfectly uni-dimensional (Wright & Linacre, 1989). This however does not obviate the necessity to avoid the exigency of measuring as opposed to counting when attempting to develop an instrument. Wright, et al (1999) points out that while no empirical process can completely account for multidimensionality scientists deal with, "…corrections for the unavoidable multidimensionality they must encounter are an integral and essential part of their experimental technique" (Wright & Stone, 1999) .

While classical sciences usually factor in adjustments for these unavoidable multidimensionalitites as an integral part of their experimental procedures it is imperative that social scientists strive to approximate the *ideal* of uni-dimensional measures if one expects to generalise the results obtained from assessments. (Wright & Masters, 1982).

Uni-dimensionality also implies *linearity*. With Rasch measurement, only characteristics thought of as linear magnitudes (i.e. weight, length, temperature, amount of education, intelligence, and strength of feeling favourable to a concept) can be described by measurement on this uni-dimensional, interval scale (Wright & Stone, 1999). In practice this would entail the allocation of the object to a point on an abstract continuum. For example, if the continuum is propensity to endorse extraverted behaviour, then the individuals may be

allocated to an abstract continuum of extraversion, one direction representing low levels of extraverted behaviour while the opposite direction represents high levels of extraverted behaviour. In essence, the concept of uni-dimensionality reflects that it is essential that the data fit the model "… in order to achieve invariant measurement within the model's uni-dimensional framework (Bond and Fox, 2007). As with the TPQ, this is one of the many reasons why individual attributes or dimensions of any complex personality assessment should be measured individually.

The Rasch model is a mathematical depiction of how fundamental measurement should function with social and psychological variables. The primary aim always being to ensure that the data conforms to the strict prescriptions of fundamental measurement – not to account for the data at hand. *Rasch fit statistics* help appraise the compromise we make between striving for uni-dimensionality and the "unavoidable exigencies of practice" (Wright & Linacre, 1992) when dealing with the idea of multidimensionality. Bond and Fox (2007) sum this up perfectly when they point out that, "In Rasch measurement, we use fit statistics to help us detect the discrepancies between the Rasch model prescriptions and the data we have collected in practice." It allows us to estimate whether each item meaningfully contributes to the measurement of a single construct by assessing the extent to which an item or person performs as expected. (Elliott, Fox, Beltyukova, Stone, Gunderson, and Zhang, 2006).

With adequate fit, a respondent with a greater level of the underlying construct (e.g. extraversion) should have the greater probability of endorsing an item of that specific construct, and similarly, one item being more difficult to endorse than another one means that for any respondent the probability of endorsing the second item is the greater level of endorsability (difficulty). (Rasch, 1960)

In essence, when $B_n > D_i$, $B_n = D_i$, and $B_n < D_i$, the possibility of endorsing extraversion is greater than 50%, equal to 50%, and less than 50% respectively. Consequently, if the item's extraversion level exactly equals the respondent's endorsement level, the probability of endorsing extraversion would be 0.5 (50%). This is the response pattern predicted by the Rasch model. Linacre's WINSTEPS uses INFIT and OUTFIT mean squares to quantify how the response patterns fit the Rasch model.

Linacre (2012) suggests that reasonable item mean-square ranges for INFIT and OUTFIT values for clinical and rating scale survey observations are 0.5 to 1.7 and 0.6 to 1.4 respectively. A mean-square of 1.0 means the measurement is accurate. When the mean-squares are lower than 1.0 we can expect the available statistical information to be less efficient and accurate. On the other hand, a mean-square higher than 1.0 starts to distort the measure and ultimately degrades the measurement system. Linacre cautions that "…mean-square values greater than 2.0…are of greatest concern (Linacre, 2012)." In measuring the fit of the TPQ measures, the clinical INFIT and OUTFIT mean-square range proposed by Linacre (2012) are used.

Separation and Reliability. Reliability generally reports the reproducibility of measures or scores. Reliability is not equivalent to accuracy or quality (Linacre, 2012) but rather an index of relative reproducibility (Linacre, 1997).

The following relationship highlights when measurement errors are independent of the measures themselves:

Reliability = True Variance / Observed Variance (1.8)

This is the reliability ratio defined by Charles Spearman in 1910. Kuder-Richardson KR-20, Cronback Alpha, and split-halves are all estimates of this ratio (Linacre, 2012).

Table 1. Sumary of 205 measured RES

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	67.5	24.0	.46	.26	1.01	-.2	1.00	-.2
P.SD	7.2	.0	.48	.02	.55	2.0	.58	1.9
S.SD	7.2	.0	.48	.02	.56	2.0	.58	1.9
MAX.	84.0	24.0	1.76	.33	4.60	8.1	4.96	8.4
MIN.	46.0	24.0	-.92	.25	.22	-4.7	.22	-4.6

REAL RMSE	.29	TRUE SD	.39	SEPARATION	1.35	RES	RELIABILITY	.65
MODEL RMSE	.26	TRUE SD	.41	SEPARATION	1.55	RES	RELIABILITY	.71
S.E. OF RES MEAN = .03								

RES RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) RES RAW SCORE "TEST" RELIABILITY = .70 SEM = 3.93

Table 1 summarises the distinctive respondent distribution extracted from WINSTEPS. These data points produce the real and model Separation and Reliability measures. Typically a value of 0.5 is accepted as the minimum meaningful reliability and 0.8 as the lowest reliability for serious decision-making (Linacre, 2012).

Also, there is a direct correlation between the reliability coefficient and the scale of the measurement error. Typically, as the standard error decreases, the Separation value increases and the Reliability measure incrementally approach its maximum of 1.0.

It is this mechanism that is applied to the TPQ measures to determine how reproducible the order of person and item measures, are.

1.2. How the Model Redefines Personality Measurement

As highlighted earlier, Georg Rasch developed a mathematical model for constructing measures. In its fundamental form, this model is based on the probabilistic relationship between an item's endorsability (difficulty) and the person's propensity to endorse (ability). The rationale behind this model is based on the premise that any difference between these two measures should determine the probability of a person either endorsing a specific extraversion item or not.

The relationship between B_n (propensity to endorse / ability) and D_i (endorsability / difficulty) is expressed as their difference (B_n - D_i). This relationship describes the probability of what happens when person *n's* endorsement level of the latent trait (extraversion in this example) is compared to item *i's* latent trait endorsability. The basic assumption being that a person with a high propensity toward extraversion, for example, has a higher probability of endorsing an item designed to measure high levels of extraversion as opposed to a person with a lower propensity toward extraversion.

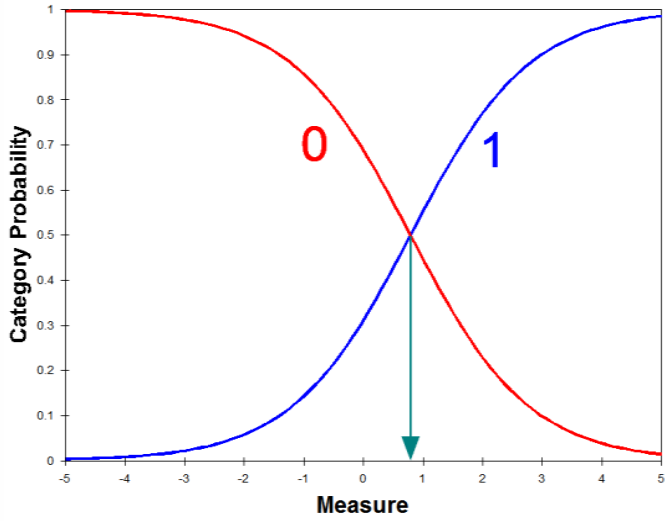
Employing the foundational model of the family of Rasch models – the dichotomous model, the aforementioned relationship (B_n - D_i) would predict the conditional probability of a binary outcome (endorsement / non-endorsement).

$$\log[P_{ni} / (1 - P_{ni})] = B_n - D_i \quad (1.9)$$

In the event of us coding endorsement as 1 and non-endorsement as 0 it is self-

evident that the probability of obtaining an endorsement (1 as opposed to 0) would be a function of the extent of the difference between the person's propensity to endorse and the endorsability of the item on that specific latent trait (B_n - D_i).

Since the ability and difficulty parameters relate to human constructs they can vary from minus infinity to plus infinity because of the variability of human nature. Because the Rasch model is a probabilistic model, and probability is restrained between zero and one, our (B_n - D_i) relationship has to comply with this rule. In order to make this happen, the (B_n - D_i) relationship is expressed as the exponent of a base **e** (a natural log function – 2.7183).



Graph 1.

In effect, when person *n* has more of the latent trait than the item *i* require, then B_n is more than D_i. This means that the propensity to endorse the underlying trait is greater than what the item requires resulting in the (B_n - D_i) relationship being positive and consequently person *n's* probability of success on the item being greater than 0.5.

So, the more person *n's* propensity to endorse the underlying trait (B_n) exceeds the item's endorsability level (D_i) the greater the positive difference and consequently the higher person *n's* probability of endorsing the latent trait.

Conversely, when the item's endorsability level D_i is much higher for person *n*, the propensity to endorse the latent trait (extraversion in this example) B_n is less than D_i and consequently the (B_n - D_i) relationship is negative resulting in person *n's* probability of endorsing the latent trait being less than 0.5.

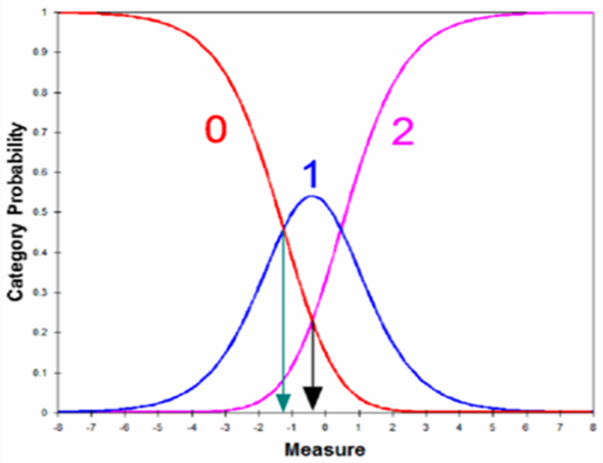
Thus, the more difficult the item is for person *n* to endorse, the greater the negative difference in the (B_n - D_i) relationship and consequently, the lower the likelihood of person n endorsing item *i*.

Georg Rasch's foundational work is based on a dichotomous (yes/no, agree/disagree, right/wrong) response by persons to items. However, when considering rating scales they are generally viewed more as "the division of the latent trait into ordered categories qualitatively advancing along a latent trait (Linacre, 2011). This expansion on Rasch's work is referred to as "polytomies"

Erling Anderson (2001) and David Andrich's (1988) early work on the Rasch model led to pivotal insights regarding Rasch polytomous analysis. Respectively, they observed that counts are sufficient for Rasch measures and that the fundamental relationship is the log-

odds of adjacent categories.

The following graph represents a 3-category Likert scale.



Graph 2.

On close examination, Graph 2 appears to be an extension of the Rasch-Andrich dichotomous model (Graph 1) with an additional parameter namely, $\mathbf{F_1}$. This $\mathbf{F_i}$ parameter is what is referred to as the "Rasch-Andrich threshold", the "step calibration" or "step difficulty" (Linacre, 2012). The following formula defines this functionality.

$$\mathbf{\log_e[P_{nij} / P_{ni(j-1)}] = B_n - D_i - F_1} \qquad (1.9)$$

The Rasch-Andrich rating scale model above specifies the probability, $\mathbf{P_{nij}}$, that person ***n*** of ability $\mathbf{B_n}$ is observed in category ***j*** of a rating scale applied to item ***i*** of difficulty $\mathbf{D_i}$ as opposed to probability $\mathbf{P_{ni(j-1)}}$ of being observed in category ***(j-1)***. Consequently, in a 3-point Likert scale, with anchors of Less, Neutral, and More, if ***j*** is "More" ***j-1*** would be "Neutral".

With reference to Graph 2, the *measure (e.g. extraversion)* axis represents the progression from "less of" the latent variable to "more of" the latent variable as one proceeds from left to right respectively. So, as one moves along the latent variable one can plot the probability of endorsing a certain level of the latent variable. At the left-hand side it looks similar to the dichotomous (Graph 1) model with a high probability of "0" and a low probability of "1".

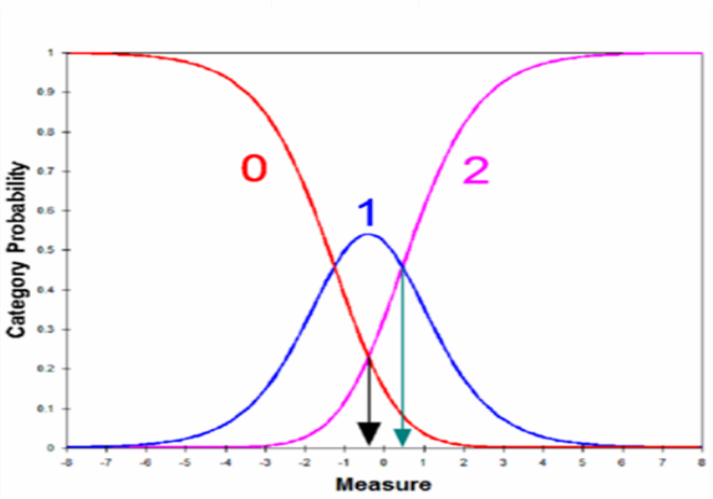
The difference now is that we reach the Rasch-Andrich threshold (green arrow) where the probability of "0" and "1" intersects – are the same. At this point, ["item endorsability" (difficulty) + "the first threshold"] = $\mathbf{D_i + F_1}$. Here, the relationship between categories 0 and 1 can be expressed as

$$\mathbf{\log_e[P_{ni1} / P_{ni0}] = B_n - D_i - F_1} \qquad (1.10)$$

On reaching the probability of "1" (blue line) there is a distinct drop in probability with an even faster drop in the "0" probability curve (red line).

In the same way, the model extends to further categories (1 and 2 in this example) as follows:

$$\mathbf{\log_e[P_{ni2} / P_{ni1}] = B_n - D_i - F_2} \qquad (1.11)$$



Graph 3.

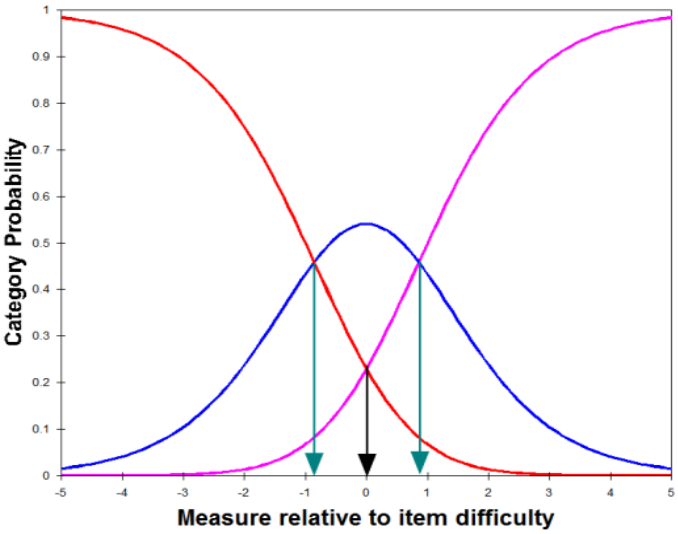
In respect to Graph 3 it is clear that the probability of "1" is higher than the probability of "2" on the left side, much like the dichotomous model. As we move up the latent variable (the right side of the graph) we reach the Rasch-Andrich threshold where curve "1" intersects with curve "2" (green arrow). At this point the probability of "1" and "2" is the same. As with "0" and "1" this intersection point is the ["item endorsability" (difficulty) + "the second threshold"] = $\mathbf{D_i + F_2}$. Again, this reflects the dichotomous model with a low probability of "1" and a high probability of "2"

Essentially, the 0-1 and 1-2 category relationships shown above are dichotomous. Given this, and the fact that the probabilities will always sum to 1, we can combine the individual relationships to construct the 3-category Likert scale in this example.

$$\log_e[P_{ni1} / P_{ni0}] = B_n - D_i - F_1 \quad (1.10)$$

$$\log_e[P_{ni2} / P_{ni1}] = B_n - D_i - F_2 \quad (1.11)$$

$$P_{ni0} + P_{ni1} + P_{ni2} = 1 \quad (1.12)$$



Graph 4.

The TPQ, like many similar instruments, use *rating scales* as an empirical medium through which the respondent can express their preferences in terms of their levels of endorsement for the construct of choice. A key consideration is always whether the data are reliable given the intention of the scale developer and the way in which the respondent in fact interpreted the scale. Given this complexity, it is important that the rating scales should not only "reflect careful consideration of the construct in question, but that they should be conveyed with categories and labels that elicit unambiguous responses" (Bond & Fox, 2007). It is therefore important to ensure that the assumptions about the quality of the measures and the utility of the rating scale be tested empirically. Once valid interval scales have been constructed, it is reasonable to proceed with Rasch statistical analysis to determine the predictive validity of a personality assessment (Wright & Linacre, 1989).

The purpose of this study is to ensure that the Townsend Personality Questionnaire complies with the "rigours of physical measures" described by Linacre (2004) by subjecting each of the five dimension to a comprehensive Rasch analysis. In this study, the author used Rasch analysis to answer the following questions:

1. *Uni-dimensionality:* To what extent do the items of each dimension of the TPQ measure a single dimension of Extraversion (Surgency), Constancy (Neuroticism), Sociability (Accommodation), Conscientiousness (Agreeableness), and Originality (Intellect or Imagination).
2. *Separation:* Can we determine the level of distinction among persons and items along each of the individual dimensions of the TPQ? How many distinct strata can be distinguished with each individual dimension of the TPQ?
3. *Reliability:* Can the internal reliability of each dimension of the TPQ be improved by managing miss-fitting items?
4. *Measurement gaps:* What measurement gaps and redundancies exist along each dimension of the TPQ personality instrument, indicating the need for adding or deleting certain types of items?
5. *Rating scale categories:* What is the optimal number of rating scale categories for each of the dimensions of the TPQ?

2. Method

2.1. Participants

Participants were 203 random respondents elicited from a structured email research request sent to a pool of 1209 individuals.

Participants comprised a very diverse age spread [21-25 (20.7%); 26-30 (16.7%); 31-35 (18.2%); 36-40 (15.3%); 41-45 (10.8%); 46-50 (4.5%); 51-55 (8.9%); 56 & Above (5.0%)], educational background [Primary School (0.0%); Secondary School (15.3%); College/Technical College (12.3%); University Degree (43.3%); Postgraduate and Professional Degree (PhD, MD, etc) (29.1%)] , and occupational categories [Administration (12.8%); Advertising, marketing and PR (5.9%); Animal and plant resources (0.5%); Arts, design, & crafts (1.5%); Construction and property management (1.5%); Counseling, social and guidance services (22.2%); Education, teaching and lecturing (8.9%); Engineering (1.0%); Finance and management consultancy (7.4%);

Healthcare (4.9%); Hospitality and events management (2.5%); Human resources and employment (7.9%); Information services (2.5%); Insurance and pensions and actuarial work (0.5%); IT, economics, statistics and management services (3.9%); Law enforcement and public protection (1.0%); Legal services (1.0%); Leisure, sport and tourism (1.5%); Logistics & transport (2.5); Publishing, media and performing arts (1.5%); Sales, retail, and buying (7.9%); Scientific services (1.0%).

Women were overrepresented and made up 76.8% of the respondents with men making up the remaining 23.2%.

Race is not included as a biographical factor since the author regards it as a fiction in terms of its utility in personality and psychological measurement and research.

2.2. Procedure

The participants responded to an email requesting their participation in the development of the TPQ personality instrument.

The cover email gave a brief overview of the nature of the questionnaire being developed as well as the theoretical rationale behind it. Additional information on the length of the questionnaire, estimated time to complete, and the confidentiality of responses were provided as well as an optional request for certain biographical data relating to gender, age, education, and occupation given the nature of the research.

The URL to the website hosting the questionnaire was included in the email. Interested participants used the link to gain direct access to the assessment on the web site hosting the TPQ. They would access the instruction page of the questionnaire directly on selecting the link.

The instructions covered the expectations of the author as well as how the participants should go about responding to the questionnaire. It also, reiterated the request in the invitation email for the participant to provide some biographical information at the end of the questionnaire.

The respondents would then use the mouse to click on a button to move to the start of the questionnaire on the next page. The participants used a mouse click to select their options on the likert scale as well as to select the demographic information at the end of the survey.

The tool restricted the respondents to one response per question and the participants had the ability to scroll up and down the questionnaire. It also allowed the participants to adjust their selections when they needed to. This was only allowed up and till the completed questionnaire was submitted.

After completing the questionnaires, the participants used a mouse to click on one of two button, "Submit" or "Back". Clicking on "Submit" sent the data to the data repository of the tool. When this occurred, the participants would receive a confirmation that the completed questionnaire had been successfully submitted. The "Back" button would allow the participants to go back into the questionnaire should they need to.

At the end of the entire data gathering process, the researcher downloaded the data for preparation and formatting for upload into WINSTEPS 3.92.1 tool for analysis.

In the light of concerns that some people may complete more than one survey online (Davis, 1999), the hosting system was set up to ensure that the responses were re-

stricted to one response from each IP address. The entire process was automated using a web-based survey tool.

Despite the typical issues associated with online delivery of assessments (control of the participants behaviour; control over motivation; inability of participants to ask questions; sample representation; manipulation and fraud; ethical problems) Musch & Reips (2000), there is general consensus that if any tool or medium is used thoughtfully and not presented as an alternate solution to traditional methods, its benefits far outweigh the possible concerns (Krantz, 2004).

Numerous studies (Myerson & Tryon (2003); Watt & Ewing (1996); Krantz (1997), et al.) affirm that when comparing web results to previously published data sets, that sample characteristics and internal consistency was the same and that the form of administration was in no way a significant influencing factor.

2.3. Instrument

The TPQ (Townsend, 2005) consists of 120 items rated on a 4-point adverb-anchored rating scale (see Table 1), ranging from 1 (very inaccurate), 2 (inaccurate), 3 (accurate), 4 (very accurate), for each dimension of the TPQ.

The TPQ is made up of five single dimensions – Extraversion, Constancy, Sociability, conscientiousness, and Originality. Each dimension comprises 24 items formatted into six facets each with anchors for the two extremes of the continuum.

The facets for extraversion are sociable, energy level, assertive, excitement seeking, unguarded, and engaging.

Table 2. Extraversion Items (n=24)

EXES4-	Hate surprises*
EXA2+	Am not afraid of providing criticism
EXEN3-	Don't enjoy being the object of jokes*
EXS2+	Don't mind being the center of attention
EXES2+	Let myself go
EXE3-	Am a very private person*
EXES1+	Am open to new experiences
EXS4-	Often feel uncomfortable around others*
EXR1+	Have a good word for everyone
EXA4-	Wait for others to lead the way*
EXR2+	Believe that others have good intentions
EXEN4-	Don't care what people think of me*
EXS3-	Only feel comfortable with friends*
EXA1+	Say what I think
EXE4-	Enjoy spending time alone*
EXE2+	Talk to a lot of different people at parties
EXEN2+	Can laugh at myself
EXEN1+	Enjoy bringing people together
EXR4-	Believe that people are essentially evil
EXS1+	Am skilled at handling social situations
EXE1+	Am ready to do battle for a cause
EXR3-	Believe that people should fend for themselves*
EXES3-	Seldom joke around*
EXA3-	Avoid challenging things*

"*" indicates item is reverse-scored

For constancy they are relaxed, emotionally stable, optimistic, impervious, self-controlled, and tempered.

Table 3. Constancy Items (n=24)

CYP3- Feel unloved*
CYE2+ Readily overcome setbacks
CYA3- Feel threatened easily*
CYA2+ Don't worry about things that have already happened
CYP4- Am concerned about the future*
CYH1+ Rarely get irritated
CYH2+ Am not easily annoyed
CYH4- Get angry easily*
CYE4- Dislike myself*
CYV1+ Am not easily bothered by things
CYV3- Am afraid that I will do the wrong thing*
CYA1+ Am relaxed most of the time
CYS1+ Never spend more than I can afford
CYV4- Feel crushed by setbacks*
CYE3- Am easily discouraged*
CYS4- Make rash decisions*
CYA4- Feel guilty when I say "no"*
CYP2+ Feel loved
CYS2+ Experience very few emotional highs and lows
CYH3- Being pleasant all the time is difficult*
CYV2+ View my mistakes as a learning opportunity
CYE1+ Feel comfortable with myself
CYP1+ Know things will turn out well
CYS3- Don't know why I do some of the things I do*
"*" indicates item is reverse-scored

For sociability they are candid, compassionate, altruistic, modest, empathic, and collaborative.

Table 4. Sociability Items (n=24)

SOAG2+ Involve others in what I'm doing
SOM4- Believe that I am better than others*
SOAG1+ Prefer to cooperate with others
SOCO2+ Accept others weaknesses
SOE4- Believe people should fend for themselves*
SOE2+ Am deeply moved by the miss-fortunes of others'
SOC1+ Am open about my views to others
SOA1+ Take an interest in other peoples' lives
SOA4- Don't like to get involved with other people's problems*
SOC3- Able to keep others at a distance*
SOE3- Believe it's the strongest that survive*
SOE1+ Anticipate the needs of others
SOAG3- Disregard the opinions of others*
SOC4- Keep my thoughts to myself*
SOCO1+ Try to forgive and forget
COA2+ Take time out for others
SOM2+ Dislike discussing personal achievements
SOM1+ Comfortable hearing another viewpoint
SOA3- Selective with the support I offer others*
SOM3- Demand to be the center of attention*
SOCO4- Believe in an eye for an eye*
SOC2+ Comfortable voicing my opinion
SOCO3- Find it hard to forgive others*
SOAG4- Act without consulting others*
"*" indicates item is reverse-scored

For conscientiousness they are self-disciplined, dutiful, competent, structured, cautious, and directed.

Table 5. Conscientiousness Items (n=24)

COC2+ Like to solve complex problems
CODI1+ Redo plans until they are perfect
COS4- Find it difficult to get down to work*
COE1+ Rarely overindulge
COD3- Rough estimates get the job done just as well*
COC4- Don't see things through*
CODI2+ Always have a back-up plan
COE3- Don't know why I do some of the things I do*
COS1+ Get chores done right away
CODI3- Deal with things as they come up*
COS2+ Accomplish my work on time
COS3- Put off unpleasant tasks*
COS3- Often forget to put things back in their proper place*
COS1+ See that rules are observed
CODI4- Things happen*
COD2+ Pay attention to details
COE2+ Need ample time before making decisions
COE4- Do things I later regret*
COS4- Prefer not planning too far ahead*
COD4- Attention to too much detail is restrictive*
COC1+ Know how to apply my knowledge
COS2+ Follow a schedule
COC3- Don't put my mind to the task at hand*
COD1+ Set high standards for myself and others
"*" indicates item is reverse-scored

For originality they are creative, complex, composure, conceptual, interest span, and conventional.

Table 6. Originality Items (n=24)

ORI2+ Enjoy thinking about things
ORF3- Am easily excited*
ORCN3- Believe there is no absolute right or wrong*
ORC2+ Love to think up new ways of doing things
ORC4- Seldom get lost in thought*
ORI1+ Ask questions that nobody else does
ORCO3- Am not interested in theoretical discussions*
ORF4- Am guided by my moods*
ORF1+ Am not easily affected by my emotions
ORC3- Do things by the book*
ORIN1+ Am good at many things
ORI4- Get bored easily*
ORCN4- Know how to get around the rules*
ORCO1+ Base my goals in life on inspiration, rather than logic
ORCN1+ Believe in the importance of tradition
ORIN4- Often feel out of my depth in conversations*
ORC1+ Get so involved with things that I forget the time
ORIN3- Can't focus on many things at the same time*
ORCO4- Try to avoid complex people*
ORI3- Feel that my interests change quickly*
ORIN2+ Comfortable talking about most topics
ORCO2+ Understand people who think differently
ORCN2+ Believe that there are universal truths
ORF2+ Always know what I'm doing
"*" indicates item is reverse-scored

Given the primacy of the assumption of uni-dimensionality in Rasch measurement (Kim & Hong, 2004), each dimension is treated as a separate and inde-

pendent construct and is measured and analysed as such.

The complete instrument comprises all 120 items randomly arranged as one continuous assessment using the Research Randomizer (http://randomizer.org).

Each dimension is coded to assist with the analysis and reporting out of each of the five dimensions. Approximately half of the items are reverse scored.

The assessment intends to measure the five domains of the Five Factor Model as described by Costa and McCrae (1992). The assessment was developed using a combination of International Personality Item Pool inventory developed by Goldberg (1999a) and items developed by the author.

3. Data Analysis

The data were analysed using the Rasch rating scale model (Andrich 1978) in WINSTEPS version 3.92.1 (Linacre, 2006). Winsteps constructs Rasch measures from simple rectangular data sets, usually of persons and items. It facilitates the analysis of dichotomous, multiple-choice, and multiple rating-scale and partial credit items.

Winsteps ensures that working of rating scales can be examined thoroughly, and rating scales can be recoded and items regrouped to share rating scales as desired.

The basic Rasch model is a dichotomous response model (Wright & Stone, 1999). It represents the conditional probability of a binary outcome of a person's propensity to endorse the underlying trait level (**B**) (respondent's ability) and an item's endorsement level on the trait (item difficulty) **(D)** (Kim & Hong, 2004):

$$\mathbf{P(x = 1) = \exp(B_n - D_i) / [1 + \exp(B_n - D_i)].} \qquad (2.1)$$

Where $\mathbf{P(x = 1)}$ is the probability of an endorsement ("yes" response to an item), $\mathbf{B_n}$ is the trait parameter of person ***n,*** and $\mathbf{D_i}$ is the difficulty of endorsing item ***i***. When $\mathbf{B_n > D_i}$, $\mathbf{B_n = D_i}$, and $\mathbf{B_n < D_i}$, the chance of a "yes" response is greater than 50%, equal to 50%, and less than 50%, respectively.

The Rasch model can be generalised to polytomous items with ordered categories. This extension of the Rasch model includes the rating scale model (Andrich, 1978) and partial credit model (Masters, 1982). As opposed to the PCM where one or more intermediate levels of endorsement might exist between complete disagreement and complete agreement, the RSM restricts the step structure to being the same for all items (Wright & Masters, 1982). The TPQ makes use of the RSM for the reason that the psychological distances between categories are the same for all items.

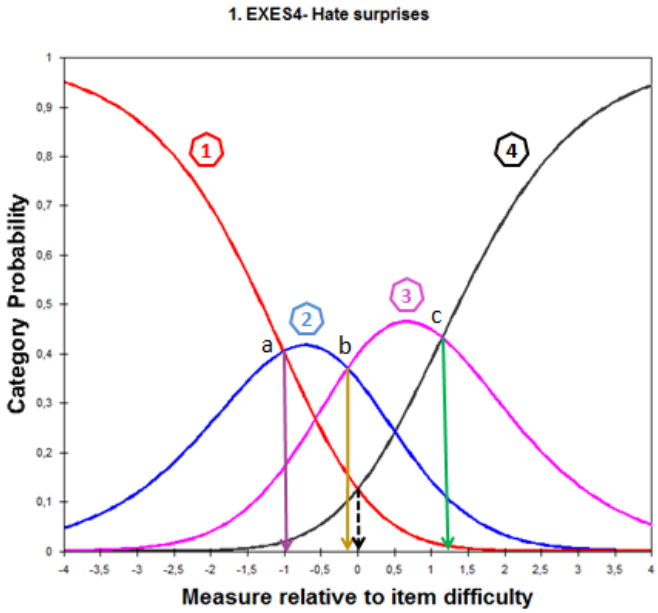
The Rasch-Andrich Rating Scale Model specifies the probability, P_{nij}, that person *n* of ability B_n is observed in category *j* of a rating scale applied to item *i* of difficulty Di as opposed to the probability $P_{ni(j-1)}$ of being observed in category *(j-1)*. So, for example, in a Likert scale (disagree, neutral, agree), j could be "agree" then *j-1* would be "neutral" (Linacre, 2012).

$$\mathbf{\log_e (P_{nij} / P_{ni(j-1)}) = B_n - D_i - F_j} \qquad (2.2)$$

$\mathbf{F_j}$ is the "Rasch-Andrich threshold" also called the "step calibration" or "step diffi-

culty"

As evident in the dichotomous Rasch model, B_n is the respondent's propensity to endorse (ability) the underlying latent trait and D_i represents the items endorsability (difficulty).



Graph 5.

As indentified by Andrich in his break-through work, we can see that a rating scale is in essence a series of Rasch dichotomies. When observing Graph 5 we have three dichotomous relationships: 1-2, 2-3, and 3-4. Given that probabilities always sum to 1, when put together one observes a 4 category picture.

$$\log_e (P_{ni2} / P_{ni1}) = B_n - D_i - F_1 \quad (2.3)$$

$$\log_e (P_{ni3} / P_{ni2}) = B_n - D_i - F_2 \quad (2.4)$$

$$\log_e (P_{ni4} / P_{ni3}) = B_n - D_i - F_3 \quad (2.5)$$

$$P_{ni1} + P_{ni2} + P_{ni3} + P_{ni4} = 1 \quad (2.6)$$

D_i (the item difficulty) represents the location where the top and bottom categories have an equally probability of being endorsed. In the case of the Rasch-Andrich model the rating scale structure takes on the same form for all the items with the different D_i values as the reference points. In the Graph 5 example, a, b, and c would be the same for all items relative to the D_i value of each item. In essence the rating scale structure slides up and down the underlying latent variable, e.g. Extraversion, for each item based on their specific item endorsability (difficulty).

Linearity and Rating scale category analysis. Lopez (1996) proposed that evaluating how respondents use the rating scale be the first step in conducting rating scale analysis.

Linacre (1997) contextualises this by pointing out that "Optimising a rating scale is

"fine-tuning" to try to squeeze the last ounce of performance out of a test. So, the first stage is to check that everything else about the test is working as well as is reasonable."

Linacre (1997) further suggests that there is no sense in trying to optimise a rating scale if the core of the assessment does not look like it works well or "…if half the sample employs a `response set'". So, only once there is a level of confidence that the core response data looks like it should work well, should the focus extend to the miss-fitting responses, ensuring that there are no data entry errors, random guessing, or other off-dimensional "bad-spots" remaining.

Only then, should the rating scale optimisation happen. Linacre (1997) also cautions that this approach is very much context driven and that careful observation at the item level should be the order of the day since "…the more you collapse categories, the more statistical and diagnostic information you lose."

Ensuring that the assessment scale is oriented with the latent variable is a fundamental prerequisite to the rating scale optimisation since it underpins the measure stability, measure accuracy (fit), sample description, and inferential qualities.

Linacre (2002) supports the idea of the importance of establishing the efficacy of the functioning of the rating scale in practice. This is based on the possibility that the respondents could react completely differently to the way the assessment designer intended (Roberts, 1994). In Rasch analysis some useful diagnostics in evaluating category usage is to examine the (a) observations of a category, (b) the observation distribution, (c) average category measures advance, (d) outfit mean squares and, (e) step difficulties advance (Linacre, 1997).

Linearity (scale orientation) analysis. As touched on previously, the underlying fundamental precept of measurement is the concept of uni-dimentionality. To this end, all measures need to be focused on one single underlying construct with each item contributing to ever increasing difficulty levels of the underlying construct. Much like measuring an individual's physical attributes that comprise a number of elements namely, weight, height, etc, one has to measure each attribute individually (Bond & Fox, 2007). Similarly, the Townsend Personality Questionnaire (Townsend, 2007) is a measure of human personality comprising five psychological dimensions. It is modeled on the Big Five personality theory and comprises five dimensions namely, Extraversion, Constancy, Sociability, Conscientiousness, and Originality. Each of these dimensions is treated as individual personality measures and are analysed as such.

WINSTEPS item polarity diagnosis was done on each of the dimensions to establish whether all their respective items were pointing in the same direction. The ensuing tables show the items ordered by their point-measure correlations and reflect whether the responses to each item are in alignment with the abilities of the respondent. The objective is to establish noticeably positive correlations (Linacre, 2012).

Table 7 summarises the relevant results from the analysis of the five di-

mensions point-measure correlations. Both Extraversion and Conscientiousness dimensions showed negatively correlated items. Respectively these were items 12, 22 and 10, 17.

Table 7. Summary of each impacted Dimension of the TPQ. Diagnostics for Original Scale miss-fitting items

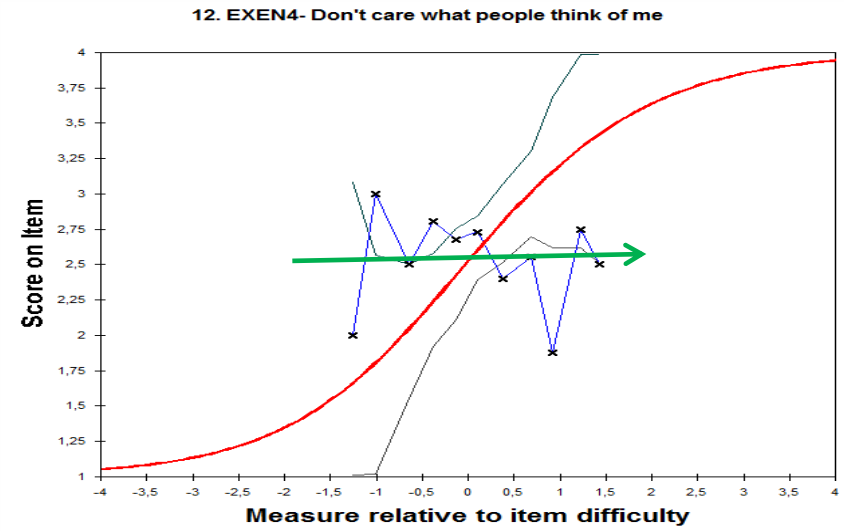
Extraversion (EX)					
Item	**Point-Measure**		**Exact Match**		**Item Description**
	CORR	**EXP**	**OBS**	**EXP**	
12	-.11	38	37.4%	42.4%	EXEN4- Don't care what people think of me
22	-.06	38	40.9%	40.8%	EXR3- Believe that people should fend for themselves
Conscientiousness (CO)					
Item	**Point-Measure**		**Exact Match**		**Item Description**
	CORR	**EXP**	**OBS**	**EXP**	
10	-.03	39	54.7%	44.6%	CODI3- Deal with things as come up
17	-.03	40	39.9%	43.5%	COE2+ Need ample time before making decisions

These four items clearly indicate that endorsement of the respective items by the respondents, contradict the direction of the extraversion and conscientiousness modeled latent variables. The expectation here is for the observed point-measure correlation to match or be as close as possible to the expected correlation (38, 38 and 39, 40 respectively). This would imply that the data matched the Rasch model.

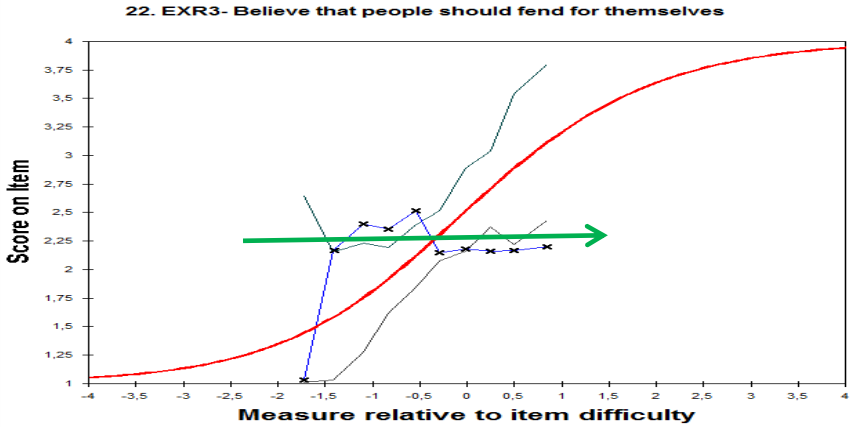
Constancy, Sociability, and Originality all indicated positive point-measure correlations. Linacre (2010) suggests that it is preferable for these correlations to be "noticeably positive". Correlations that are closer to zero or negative are more than likely counterintuitive to the underlying direction of measurement. This may also be an indication of response problems to reverse-coded items or ambiguity based on the item structure.

In order to further evaluate the aspect of uni-dimensionality of the TPQ measures each of the negatively correlated point-measures are further investigated.

On analysis of Graph 6 and Graph 7 representing the extraversion negative point-measure correlation items and Graph 8 and 9 measuring the negative point measure items of the conscientiousness dimension it is clear that the items are being responded to in unexpected ways. The red lines represent the item characteristic curve as anticipated by the Rasch model. The turquoise lines represent the 95% confidence band that is 1.96 standard errors vertically away from the red Rasch model line. The blue line is the empirical Item Characteristic Curve (ICC). The "x"'s along this line represent the respondents with measures close to the measure of "x" on the x-axis.

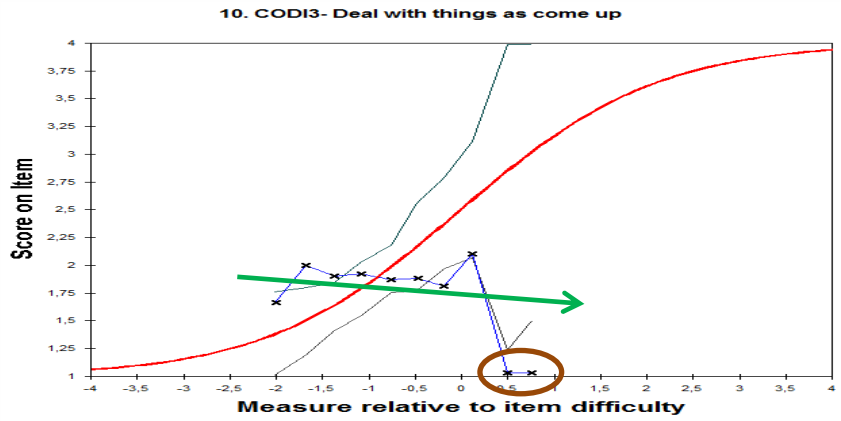


Graph 6. Extraversion Linearity Flags

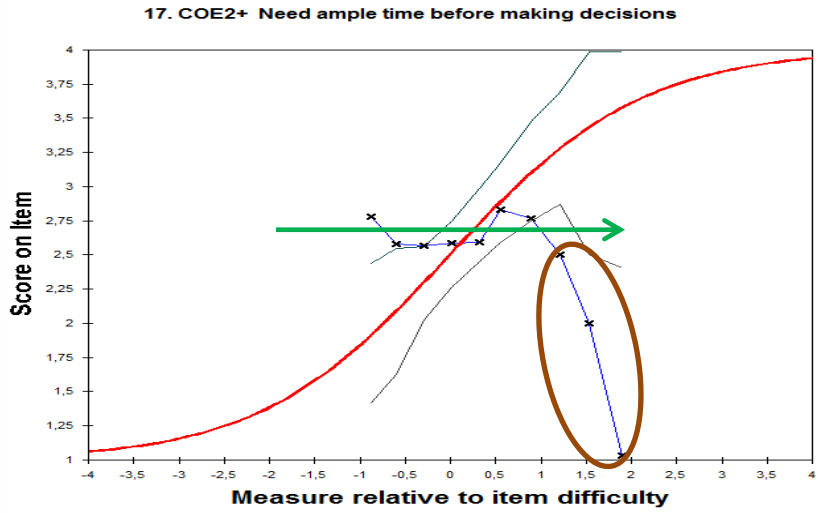


Graph 7. Extraversion Linearity Flags

The aim here is for the blue empirical ICC line to approximate the red Rasch model line as closely as possible (Linacre, 2014).



Graph 8. Conscientiousness Linearity Flags



Graph 9. Conscientiousness Linearity Flags

Graph 6 and 7 shows a clear deviation from the model with a number of responses outside the confidence interval lines. The green arrow shows the distinctly unexpected direction of the responses to the modeled ICC. These response patterns clearly indicated that expectations of the more extraverted respondents endorsing extraversion were not evident.

These items clearly reflected either misinterpretation of the reverse-coded item and or ambiguity in the way the items were constructed. In any event, both "Don't care what people think of me" and "Believe that people should fend for themselves" show strong evidence for deletion as they do not contribute to the underlying constructs.

Graphs 8 and 9 similarly show deviation from the modeled conscientious ICC. Interestingly though, both items appear to comprise two distinct parts. The green arrows obviously ignoring the red modeled ICC while some respondents circled in brown are clearly indicating a completely different narrative. Linacre (2012) suggests that "All items must be about the same thing, our intended variable, but then be as different as possible, so they tell us different things about the latent variable". Both "Deal with things as they come up" and "Need ample time before making decisions" show strong evidence that they do not contribute to the underlying measure. Since they possibly speak to a different or secondary dimension they display strong evidence for deletion.

Person and Item Separation and Reliability. Table 8 summarises the changes in Person and Item Separation and Reliability by removing miss-fitting Items and or Respondents. The analysis objective is to obtain optimal Separation, Reliability, INFIT and OUTFIT mean square measures by minimising or completely doing away with all miss-fitting items and / or respondents.

The Person and Item Separation statistic provide a means of establishing "the number of statistically different levels of person ability that are distinguished by the items" (Elliot et al, 2006). Wright and Masters (1982) suggest a minimum separation value of 2.0. In essence this statistic defines a "hierarchy of items along the measured variable (Kim et al, 2004).

Reliability refers to the reproducibility of the measures. Linacre (2012) suggests that a score of 0.5 borders on marginal reliability while a score of 0.8 (a separation of approximately 2.0) should be regarded as the lowest cut-off for serious decision-making given the ceiling of 1.0.

Regarding miss-fitting respondents, Keeves and Masters (1999) suggest that often miss-fitting respondents reflect individuals at the "extremities of the trait distributions". They agree that where the propensity to endorse an item (B_i) – the endorsability (D_i) of the item is greater than 2.0 and, they miss-fit the Rasch model, that "...these cases should be removed from analyses" (Curtis, 2004).

Since increasing the sample size will not generally impact the person reliability scores, unless they contribute a wider range of ability (Linacre, 2012), the main focus of the TPQ measurement improvement is centered on the removal of miss-fitting items.

Referencing Table 8, we can observe that Constancy, Sociability, and Originality TPQ dimensions show great item separation and reliability. Respectively, 7.64 and 0.88, 6.42 and 0.98, 7.75 and 0.98. Given there were no miss-fitting items in these three dimensions, the miss-fitting respondents were deleted producing improved scores of 7.97 and 0.98, 6.87 and 0.98, and 8.06 and 0.98.

Table 8. Summary of Changes in Person and Item Separation and Reliability by Removing Miss-fitting Items and / or Respondents

Analysis	Separation (G)		Reliability		In-fit mean square	Out-fit mean square	Number of Miss-fitting items	Number of Miss-fitting respondents
Extraversion	Respondent	Item	Respondent	Item				
Original	1.35	7.51	.65	.98	1.01	1.00	2 (12,22)	4 (104,198,31,86)
Adjusted (Items)	1.59	7.65	.72	.98	1.01	1.01	0	4 (104,198,31,86)
Adjusted (Respondents)	1.38	7.84	.65	.98	1.00	1.00	2 (12,22)	0
Adjusted (Items & Respondents)	1.60	7.97	.72	.98	1.00	1.01	0	0
Constancy								
Original	2.38	7.64	.85	.88	1.00	1.02	0	5 (104,63,198,159,6)
Adjusted (Items)								
Adjusted (Respondents)	2.39	7.97	.85	.98	.99	1.01	0	0
Sociability								
Original	1.37	6.42	.65	.98	1.00	1.00	0	7 (104,198,31,4,137,141,6)
Adjusted (Items)								
Adjusted (Respondents)	1.42	6.87	.67	.98	1.00	1.00	0	0
Conscientiousness								
Original	1.56	6.90	.71	.98	1.00	1.00	2 (10,17)	6 (104,198,85,109,86,6)
Adjusted (Items)	1.76	6.44	.76	.98	1.00	1.00	0	6 (104,198,85,109,86,6)
Adjusted (Respondents)	1.60	7.28	.72	.98	.99	1.00	1 (17)	0
Adjusted (Items & Respondents)	1.79	6.78	.76	.98	.99	1.00	0	2 (18,150)
Originality								
Original	1.39	7.75	.66	.98	1.00	1.00	0	7 (43,12,102,198,169,7,6)
Adjusted (Items)								
Adjusted (Respondents)	1.37	8.06	.65	.98	.99	1.00	0	0

Extraversion and Conscientiousness show good separation and reliability of 7.51 and 0.98, 6.90 and 0.98 respectively. They however both have miss-fitting items namely 12 and 22, 10 and 17 respectively. Removing the miss-fitting items from both dimensions resulted in readings of 7.65 and 0.98, 6.44 and 0.98 respectively. Further, treating the miss-fitting respondents as excluded produced even better readings of 7.84 and 0.98, 7.28 and 0.98 respectively. These responses were omitted since they were negatively correlated to the underlying measure and did not comply with the expected objective ordering along the measures (Wright & Masters, 1982). This indicates either a certain level of misun-

derstanding of the items intention, its reverse-order structure, or simply just random responses.

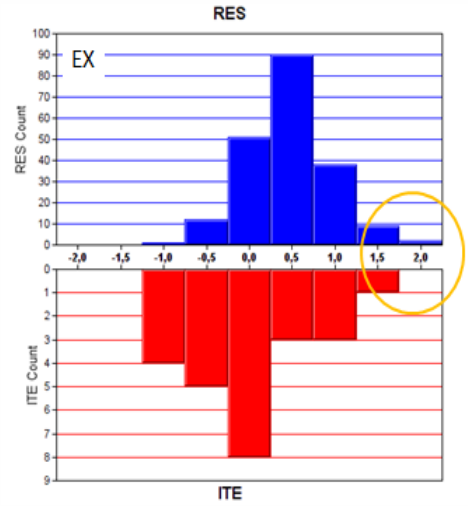
Finally, removing both miss-fitting items and respondents resulted in readings of 7.97 and 0.98, 6.78 and 0.98 respectively. In both extraversion and conscientiousness, we observed a marked improvement when removing both miss-fitting items as well as negatively correlated respondents. The post-analysis item reliabilities for Extraversion, Constancy, Sociability, Conscientiousness and Originality are all exactly 0.98. According to Linacre (2010) this is well above the minimum reliability threshold of 0.80 for "serious decision-making".

Further, Rasch proposes chi-square fit statistics be used to ascertain how effectively a set of data fit the requirement of his model (Bond & Fox, 2007). These fit statistics are presented as INFIT and OUTFIT mean squares statistics in programs like WINSTEPS (Linacre, 2006; Wright, 1984; Wright & Masters, 1981).

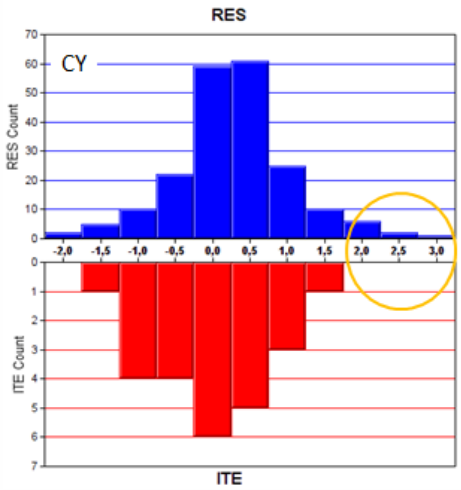
Infit and outfit measures are reported as ratio scales with "…an expected value of +1 and a range of 0 to positive infinity" (Bond & Fox, 2007). Since infit and outfit values are always positive a value of, for example, 1.25 is indicative of 25% more variation in the observed data than the expected model. Equally, an infit or outfit value of, for example, 0.82 is indicative of 18% less variation than its modeled expectation of 1.0. Bond & Fox (2007) cautions that "There are no hard and fast rules" when interpreting Rasch fit statistics however, there are some reasonable guidelines namely 0.8-1.2 for high stakes multiple-choice tests, 0.7-1.3 for standard multiple-choice tests, 0.6-1.4 for Likert rating scales, 0.5-1.7 for clinical observations, and 0.4-1.2 for Judged tests. Essentially, an infit score above 1.0 is regarded as underfitting the modeled data thus presenting "too much unpredictability" while an outfit less than 1.0 indicates that the data in fact overfits the model implying that endorsements on various items are much too predictable.

Based on these data, all further analysis is done after removing miss-fitting items 12 and 22, respondents 104, 198, 31, and 86 from the Extraversion measure and miss-fitting items 10 and 17, respondents 104, 198, 85, 109, 86, 6 from the Conscientiousness measure.

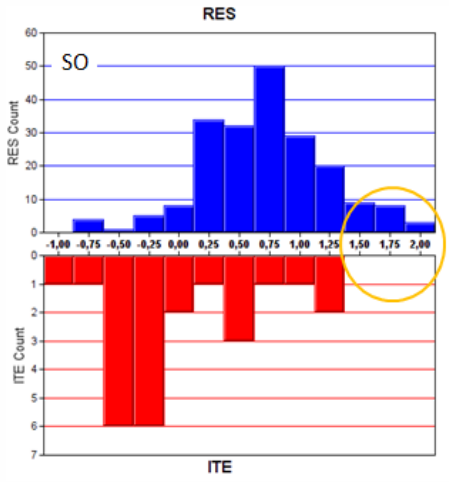
Sampling and Measurement Gaps. The following person-item distribution maps (Graphs 10 to 14) show the respondents in blue and the items in red. In line with the basic Rasch idea of *parameter separation* (Bond & Fox, 2007) we observe both the respondents as well as the items being measure on the same logistic scale. Given these measures share the same scale it is easy to establish the relative order of endorsability levels of the items in relation to each other as well as how they spread across the same scale the order of respondents are in relation to their propensity to endorse levels.



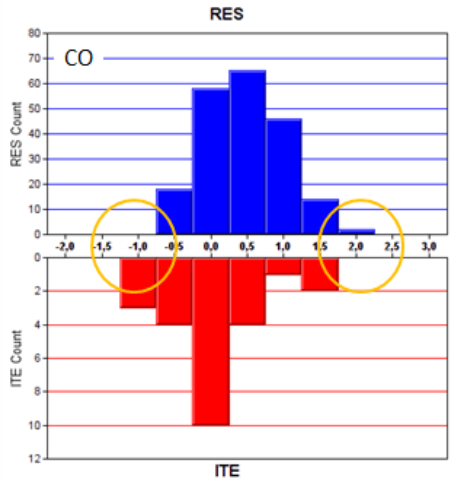
Graph 10. Extraversion



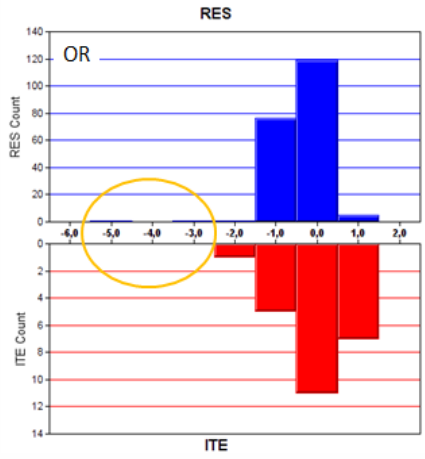
Graph 11. Constancy



Graph 12. Sociability



Graph 13. Conscientiousness



Graph 14. Originality

The logistic nature of the Rasch model enables one to visualise the levels of strength or endorsability of the items relative to one another as well as where each respondent lies on the exact same scale without any influence one on the other.

The yellow circles in Graphs 10 to 14 display either under-sampled or low scoring respondents and / or too limited or expansive item structure of each TPQ dimension. Graph 10 has a few respondents that recorded 2.0 and above logits on the extraversion Endorsability scale with no equivalent items of that strength. Graph 11 shows a similar pattern as Graph 10 but with some respondents displaying a lower propensity to endorse Constancy than there are items to measure at the level of below -2.0 logits. Graph 12 shows a number of respondents displaying a propensity to strongly endorse Sociability at and above 1.5 logits and well up to 2.0 logits. The item measures span -1.0 to 1.25 logits. Graph 13 shows a few respondents with high propensity (2.5 logits) to endorse Conscientiousness while the items stop at approximately 1.5 logits. Finally, Graph 14 shows a small number of respondents with very low (almost off the scale) propensity to endorse Originality while the items seem to represent a reasonable spread of items as a measure.

Rating scale category analysis. The average measures across categories are an empirical indicator of the context in which the rating scale categories are used. Because higher categories are intended to reflect higher measures, the average measures across categories are expected to increase monotonically (Kim & Hong, 2004). Similarly, advancing threshold difficulties imply that each category in turn is most likely to be chosen. Disordered step difficulties (thresholds) suggest that a category may not be observed as one advances along the variable. Linacre (1997) cautions that, "Disordered step difficulties do not mean that the categories are out of order". Consequently, any "…decision to eliminate or combine narrow categories must be decided substantively based on the reasons for selecting the rating categories."

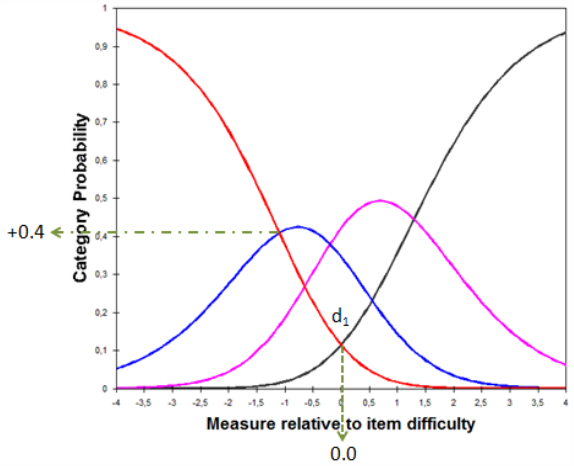
For the 4-point scales of each of the 24-item dimensions of the TPQ (extraversion (EX), constancy (CY), sociability (SO), conscientiousness (CO, and originality (OR)), the *average measures* increased with the category label (**EX**: -0.45, 0.01, 0,57, 1.05, 0.94; **CY**: -0.86, -0.18, 0.48, 1.23; **SO**: -0.14, -0.16, 0.77, 1.23; **CO**: -0.32, -0.01, 0.60, 1.10; **OR**: -0.34, -0.07, 0.39, 0.85) for categories 1 to 4, respectively (Table 7). This suggests that the rating scale categorisation is satisfactory.

In addition, the *thresholds estimates* across all dimensions were ordered, with logits of **EX**: -1.01, -0.14, 1,15; **CY**: -1.16, -0.12, 1.28; **SO**: -1.21, -0.12, 1.33; **CO**: -1.19, -0.04, 1.23; **OR**: -0.87, -0.17, 1.05) respectively (Table 7).

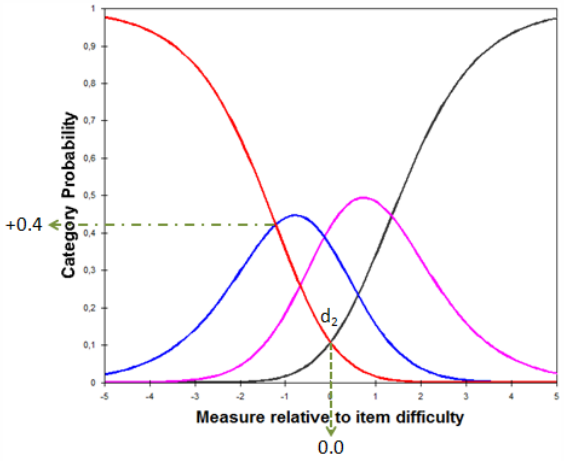
Given that all the thresholds are increasing monotonically and are reasonably spaced from one another, it can be assumed that there is a clear and meaningful progression along each of the five variables. Linacre (1995) suggests this separation should be at least 1.4 logits apart but no greater than 5 logits.

In addition, the probability category curves in Graphs 15 to 19 clearly show distinct response categories. These represent the probability of response categories as a function of the respective underlying trait. As described previously, each intersecting point of the adjacent rating scale category is regarded as the estimated threshold value of the higher of the two categories. All probability curves clearly do not overlap the adjacent categories excessively implying that they provide enough clarity to determine a specific point along each of the respective vari-

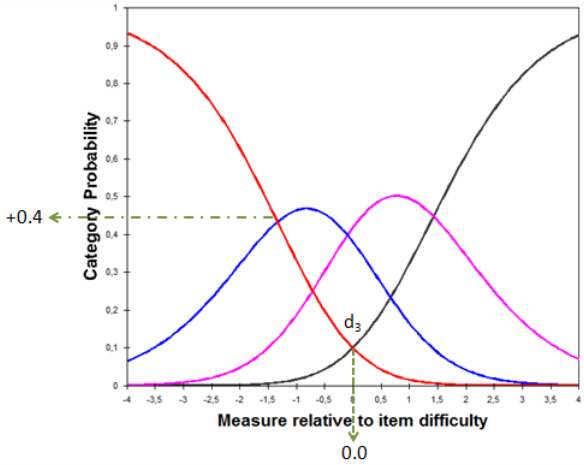
ables.



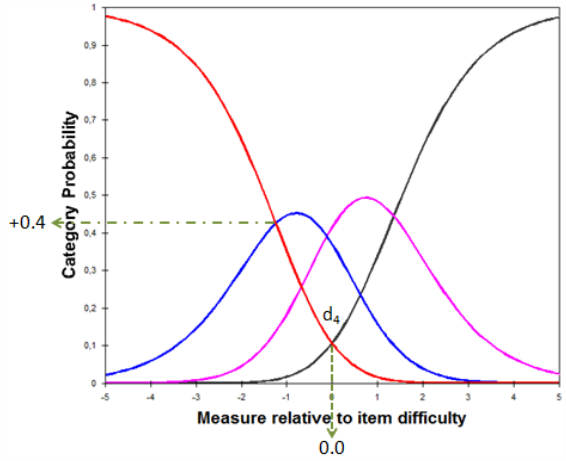
Graph 15. Extraversion



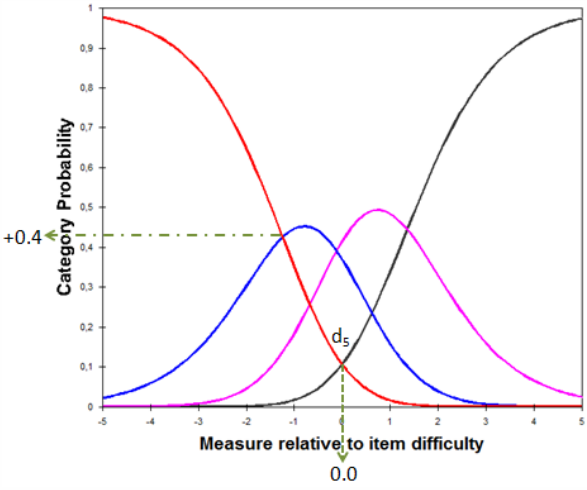
Graph 16. Constancy



Graph 17. Sociability



Graph 18. Conscientiousness



Graph 19. Originality

For example, if a respondent's propensity to endorse "Am relaxed most of the time" is 2 logits higher than the endorsability level of this specific item, the expectation would be that the respondent would probably endorse it at level 3.

Table 9. Summary of each Dimension of the TPQ Five-Point Rating Scale Category Functioning (1234). Diagnostics for Original Scale 1234

Category label Extraversion (EX)	Observed count	Average measure	In-fit mean square	Out-fit mean square	Threshold calibration
1 (very inaccurate)	531	-.45	1.06	1.09	NONE
2 (inaccurate)	1173	.01	.92	.90	-1.01
3 (accurate)	1852	.57	.90	.88	-.14
4 (very accurate)	1316	1.05	1.04	1.06	1.15
Constancy (CY)					
1 (very inaccurate)	690	-.86	1.03	1.08	None
2 (inaccurate)	1315	-.18	.94	.92	-1.16
3 (accurate)	1742	.48	.92	.89	-.12
4 (very accurate)	1125	1.23	1.06	1.12	1.28
Sociability (SO)					
1 (very inaccurate)	330	-.14	1.16	1.18	None
2 (inaccurate)	1068	.16	.89	.88	-1.21
3 (accurate)	2021	.77	.92	.89	-.12
4 (very accurate)	1453	1.23	1.02	1.03	1.33
Conscientiousness (CO)					
1 (very inaccurate)	471	-.32	1.15	1.18	None
2 (inaccurate)	1283	-.01	.87	.83	-1.19
3 (accurate)	1862	.60	.86	.84	-.04
4 (very accurate)	1256	1.10	1.03	1.05	1.23
Originality (OR)					
1 (very inaccurate)	682	-.34	1.01	1.02	None
2 (inaccurate)	1277	-.07	.93	.92	-.87
3 (accurate)	1768	.39	.95	.97	-.17
4 (very accurate)	1145	.85	1.04	1.04	1.05

Note: *Observed count* include all respondents' responses for each category. The *Average measure* reports the average propensity to endorse the underlying latent trait of the respondents who selected the response. Typically higher categories are expected to be endorsed by respondents with a higher propensity to endorse the underlying trait. Reasonable Item Mean-square Ranges for INFIT and OUT-FIT for Likert Rating Scale observations is between 0.6 – 1.4 (Bond & Fox, 2007).

The evidence of ordered increases across all dimensions in Table 9 is an indication that the average measures are ordered. In summary, this is a clear indication that the rating scales, across all dimensions, are being used as the author intended.

In summary, this is a clear indication that the rating scales, across all dimensions, are being used as the author intended.

4. Results

4.1. Diagnostic Analysis

Linearity analysis. After assessment of all five dimensions of the TPQ namely, Extraversion, Constancy, Sociability, Conscientiousness, and Originality, the fit statistics and item point-measure correlations showed that only two of the dimensions (Extraversion and Conscientiousness) had two items respectively that were negatively correlated to their underlying measures. The analysis to determine uni-dimentionality of these dimensions entailed computing point-measure correlations as well as infit and outfit mean-squares using the WINSTEPS version 3.92.1 (Linacre, 2006) software.

Point-measure correlations are expected to be positive if the items are comply-

ing with the underlying Rasch models expectation. Negative correlations are a clear indication that the items are *not* behaving as expected.

In addition, the mean-square fit statistics has reasonable range expectations of 0.6 to 1.4 for Likert-styled behavioural surveys (Linacre, 2012). It is also suggested that the higher (>1.0) the mean-square values are the less homogenous the underlying construct is in regard to the other items in the measure while, low (<1.0) items typically point to levels of redundancy among items (Kim, et al, 2004).

The two negatively correlated items for Extraversion and Conscientiousness were -0.11, -0.06 and -0.03, -0.03 respectively while their infit and outfit mean-square values were 1.35, 1.39 and 1.22, 1.25 respectively. All the mean–square values pointed to parameter-level mean-square fit statistics that are productive for measurement. Given that the mean-square values are way below the critical 2.0 level (Linacre, 2012) and within the 0.6 to 1.4 range (Linacre, 2012), the negative point-measure correlations were used to determine that these items were to be deleted.

To further refine the linearity requirement of the underlying constructs of the five dimensions, all negative point-measure correlation respondent values were deleted. This was done taking into account Bond & Fox's (2007) caution to use the fit statistics to identify problematic items and person responses instead of just for the removal of items. As shown in Table 8, deletion of all miss-fitting items improved the underlying measures.

All further Rasch analysis was performed with 22 Extraversion items, 24 Constancy Items, 24 Sociability items, 22 Conscientiousness items and 24 Originality items. In addition, Extraversion respondents 104, 198, 31, 86, Constancy respondents 104, 63, 198, 159, 6, Sociability respondents 104, 198, 31, 4, 137, 141, 6, Conscientiousness respondents 104, 198, 85, 109, 86, 6, and Originality respondents 43, 12, 102, 198, 169, 7, 6 were treated as missing data.

Separation and reliability analysis. Table 8 shows the Rasch respondent and item separation statistics for each measure underlying the TPQ namely, Extraversion, Constancy, Sociability, Conscientiousness, and Originality (G = 1.60 and 7.97, 2.39 and 7.97, 1.42 and 6.87, 1.79 and 6.78, 1.37 and 8.06 respectively and Reliability = 0.72 and 0.98, 0.85 and 0.98, 0.67 and 0 .98, 0.76 and 0.98, 0.65 and 0.98 respectively.

So, firstly, for Extraversion, the person separation is 1.60, corresponding to a person reliability of 0.72 with the item separation of 7.90 corresponding to a test reliability of 0.98. This indicates that this measure can distinguish between respondents with high and low propensities to endorse Extraversion across 1.6 performance levels.

Secondly, for Constancy, the person separation is 2.39, corresponding to a person reliability of 0.85 with the item separation of 7.97 corresponding to a test reliability of 0.98. This indicates that this measure can distinguish between respondents with high and low propensities to endorse Constancy across 2.39 performance levels.

Thirdly, for Sociability, the person separation is 1.42, corresponding to a person reliability of 0.67 with the item separation of 6.87 corresponding to a test reliability of 0.98. This indicates that this measure can distinguish between respondents with high and low propensities to endorse Sociability across 1.42 performance levels.

Fourthly, for Conscientiousness, the person separation is 1.79, corresponding to a person reliability of 0.76 with the item separation of 6.78 corresponding to a test

reliability of 0.98. This indicates that this measure can distinguish between respondents with high and low propensities to endorse Conscientiousness across 1.79 performance levels.

Finally, for Originality, the person separation is 1.37, corresponding to a person reliability of 0.65 with the item separation of 8.06 corresponding to a test reliability of .98. This indicates that this measure can distinguish between respondents with high and low propensities to endorse Originality across 1.37 performance levels.

Examination of the probability curves (Graphs 15 to 19) revealed, in all five dimensions, that all categories increased monotonically across each rating scale and that each was always the most probable for a specific part of the underlying continuum (Bond & Fox, 2007).

Sampling and Measurement Gaps. Graphs 10 to 13 show the spread of respondents and their levels of ability to endorse the relevant underlying constructs. Similarly, and on the same scale, they show the linear layout of the various items underlying each construct. Extraversion, Constancy, and Sociability appears to have a few respondents in the sample that consistently endorse the *very accurate* category while the underlying measure shows fewer items of that strength available to measure at those levels. While Conscientiousness displays the same aforementioned pattern, it also has no respondents in the sample endorsing the *very inaccurate* category while the item scale has items measuring the underlying trait at that specific level. Originality, on the other hand, has a balanced respondent sample spread and representative items at the category levels from "*inaccurate*" to *"very inaccurate"*.

However, toward the *"very inaccurate"* side of the scale there are a few extreme endorsing respondents with no equivalently weighted items to measure at that level of the underlying measure. While the spread of items are reasonable across all five measures, adding some more difficult to endorse items to the Extraversion, Constancy, Sociability, and Conscientiousness measures would assist in capturing the respondents endorsing these dimensions at the higher end of the scale. Originality, on the other hand could do with a few items measuring the lower end of this measure. The respondents that represent at the extreme end of the low scale of Originality do raise a question of fit to the underlying model since there appears to be a lot of noise in those data points.

These extreme respondents represent 2.5% of the total number of respondents with two showing clear negative point-measure correlations while two show close-to-zero correlations (Linacre, 2012). Given the fact that a marginal amount of respondents are producing the extreme low endorsements, caution should be applied when making any decision to include any additional items to improve measurement at the lower Originality levels (Graph 14).

On examination of the observed frequencies of each rating category of each measure's items, the sample size and spread was adequate according to Rasch's expectation. Table 7 clearly show that the expectation of 10 observations per response category is overwhelmingly met across all TPQ measures (Linacre, 2002).

4.2. Category Analysis

Extraversion, Constancy, Sociability, Conscientiousness, and Originality in Graphs 15 to 19 show the analysis of the TPQ four-point rating scale categories for

each measure. Here the probability of the response categories is expressed as a function of the respondent's endorsement level of the dimension. Also, the probability of endorsing a specific category is the likelihood of endorsing a given rating scale category at that level of Extraversion, for example. In this context, the intersection of the adjacent rating scale categories can be seen at the estimated threshold value of the higher of the two categories. For example, the threshold value for Category 1 is −1.01 (reported in Table 8 and visually represented in this Graph 15). Consequently, the probability of choosing Category 1 at this level is marginally above .4, as shown by the height of the intersection on the *y* axis and approximately -1.3 on the *x* axis. These intersections reflect the points on the scales where the probability of selecting, in this example, either Category 1 or Category 2 is equally probable. In addition, the probability curves in Graph 15 to 19 are centered on the scale value $d_{1\text{-}5} = 0.0$ logits.

These results all display monotonically increasing thresholds for all five TPQ measures and relatively equal distances between the categories (Table 8). Failure of threshold parameters to increase monotonically is regarded as "step disordering" (Linacre, 2002) which in turn results in low probabilities of observing all categories in the scale. In addition, and as highlighted earlier, all TPQ measures displayed person separation values greater than 0.80 (1.60, 2.39, 1.42, 1.79, and 1.37 respectively) which is regarded as great reliability for measurement purposes (Fox & Jones, 1988). As a consequence of these results, it was established that the four-category likert scale (1 – very inaccurate, 2 – inaccurate, 3 – accurate, 4 – very accurate) is the most appropriate for the Townsend Personality Questionnaire (TPQ). Graphs 15 to 19 were generated with WINSTEPS 3.92.1 (Linacre & Wright, 2006).

4.3. Improving the Townsend Personality Questionnaire (TPQ)

On initial analysis of each of the five TPQ dimensions it became clear that Items 12 and 22 of the Extraversion dimension and Items 10 and 17 of the Conscientiousness dimension showed counterintuitive point-measure correlations. These results clearly indicated that the responses did *not* align with the respondent's propensity to endorse (ability) the underlying trait. These results are contrary to Rasch's expectation that higher respondent measures should result in higher endorsement of items and similarly, higher endorsement of items should equate to higher respondent propensity to endorse (Linacre, 2012).

Table 7 shows EXEN4- Don't care what people think of me at -0.11, EXR3- Believe that people should fend for themselves at -0.06, CODI3- Deal with things as they come up at -0.03 and

COE2+ Need ample time before making decisions at -.03. Given that the items are noticeably negative and a mixture of positively and negatively worded items, it is clear that simply rescoring the items will not resolve these conflicts.

In a future iteration of the TPQ possible reframing of EXEN4-, EXR3-, CODI3-, and COE2+ should remove the confusion and improve the approximation of the Extraversion and Conscientiousness point-measure values of 38, 38, 39, and 49 respectively.

Overall, after removing the miss-fitting items, all the remaining items across the five personality measures show clear linearity and no redundancy.

5. Discussion

The purpose of this study was to evaluate the Townsend Personality Questionnaire (TPQ) using Rasch analysis in order to establish the efficacy of the underlying measures comprising the TPQ namely Extraversion, Constancy, Sociability, Conscientiousness and Originality. As far as has been established the TPQ is the first Big-Five assessment that has been developed using Rasch methodology.

The TPQ is developed for both personal as well as workplace development of people. Unlike most assessments that are based on theories of personality that have resulted from one particular psychologist's theory and opinion about human nature, the TPQ is based on the Five-Factor model - a concept that it is founded on the idea that five main factors are necessary and sufficient for broadly describing human personality. The five factor theory is among the newest models developed for describing personality and has demonstrated that it is among the most practical and applicable models available in the field of personality psychology (Digman, 1990).

Also, often referred to as the 'Big Five' (Ewen, 1998, p.140), this model represents the most widely acknowledged general model of the structure of personality (Betram & Brown, 2005). It incorporates five different variables into a conceptual model for describing personality (Popkins). For this reason Howard & Howard (2004) point out that "…it is from language itself, and not theories, that we must extract the source metaphor for describing personality".

So, in the TPQ, the data are grouped into five Personality Dimensions (Extraversion (EX), Constancy (CY), Sociability (SO), Conscientiousness (CO), Originality (OR)) and thirty facets. These dimensions represent the most common behavioural styles exhibited by people in general and are comprised of the typical behaviours (facets) constituting each dimension.

For this reason, and unlike traditional measure construction, the TPQ has five distinct measures that collectively and conceptually comprise the personality measure. Traditionally, instead of focusing on constructing *measures* of the human state, psychologists and social scientists inadvertently applied sophisticated statistical procedures to nothing more than counts of observed events or levels of performance (Bond and Fox, 2007). So, despite having a measure of utility over the last 100 years, the traditional approaches to construction and evaluation of measures are proved not to be unproblematic and error-free (Elliott, Fox, Beltyukova, Stone, Gunderson, and Zhang, 2006). In Fisher's (2002) words, '…if we can't generalize from our data, no amount of statistical *hocus pocus* is going to construct meaningful results.'

The TPQ therefore uses Rasch methodology as its primary point of departure during the instrument development process since "… for any measurement to be meaningful, it must be based on the "arithmetical properties of the interval scales used" (Wright & Linacre, 1989).

In order to deal with some of the aforementioned inadequacies highlighted above and to provide a sufficient foundation for traditional methods Georg Rasch (1960, 1980) developed a revolutionary model for measurement in the social sciences. These Rasch models form the framework within which assessment developers can evaluate the utility of their measures (Elliott, Fox, Beltyukova, Stone, Gunderson, and Zhang, 2006).

This study further proposes that future behavioural research should prioritise

the application of Rasch methodology as the only method able to "transform raw data from the human sciences into abstract equal-interval scales". (Bond & Fox, 2001). It is a logistic item response model that independently scales both items and persons along the same underlying construct (Kahler, et. al., 2004).

In reference to the specific findings about the analysis of the TPQ, all items constituting the five dimensions show overall positive point-measure correlations with the exception of EXEN4- Don't care what people think of me (reverse worded) and EXR3- Believe that people should fend for themselves (reverse worded) items in the Extraversion construct and CODI3- Deal with things as they come up (reverse worded) and COE2+ Need ample time before making decisions items in the Conscientiousness construct. All four items were negatively correlated to their respective underlying measures indicating that the respondent's endorsement of these items contradicted the direction of the aforementioned latent variables respectively. These items were removed to facilitate further Rasch analysis of the five measures. The results reflected a considerable improvement once items 12, 22, 10, and 17 were removed (Table 8). For a future iteration of the TPQ the following adjustments are proposed for the negatively correlated items namely, EXEN4+ Care what people think of me (*positively framed*), EXR3- Believe that people *should be self-sufficient* (reverse worded), CODI3+ Deal with things *immediately (positively framed)* and COE2+ *Apply my mind* before making decisions. On re-evaluating the measures after the removal of the miss-fitting items, the expected correlations all approximated the underlying five variables. Linacre (2012) suggests that it is pointless proceeding with further analysis if the underlying items of each variable are not functioning as they should be.

On assessing the adequacy of the four-scale category functioning for Extraversion, Constancy, Sociability, Conscientiousness, and Originality, the results suggested that all dimensions of the TPQ had rating categories that advanced monotonically from *very inaccurate* to *very accurate*. In addition, respondents appeared to effectively discriminate between Category 1 (very inaccurate), Category 2 (inaccurate), Category 3 (accurate), and Category 4(very accurate). Table 8 and Graphs 15 to 19 numerically and graphically show this equally spaced and monotonic linear spread of categories. Given all the thresholds of each dimension are monotonically ordered, it does not violate the principles underlying the Rasch model and for this reason they are retained as the measurement scale structure for the ^{TPQ} (Andrich, 1978).

Graphs 15 to 19 focus on the relative distributions of item endorsability (difficulty) and respondent's propensity to endorse (ability) estimates for Extraversion, Constancy, Sociability, Conscientiousness, and Originality.

All the distribution maps appear to have functioned effectively in that each of the five measures were represented well across the full range of respondents' trait abilities. Even after the removal of Extraversion items 12, 22 and Conscientiousness items 10, 17, all measures displayed sufficient overlap of respondents propensity to endorse the underlying trait and the item endorsability (difficulty) levels. There was evidence of some respondents endorsing slightly above the strongest items on each of the measures. However, each set of items reasonably measure respondents with both high and low Extraversion, Constancy, Sociability, Conscientiousness and Originality traits. The marginal number of respondents who were not accommodated with items of equal strength at the upper end of each measure is not sufficient to conclude that the instru-

ment may be prone to a "ceiling effect" such that, it may not effectively "...detect the full variation in a population" (Kim & Hong, 2004). However, there is no reason why a few stronger items across each measure should not be considered.

Given the variability of human behavioural measurement there were limitations in this specific study. First, the sample was skewed from a gender perspective. While the automated assessment links were sent to a random sample of 1209 individuals, of the 203 that responded, women were overrepresented and made up 76.8% of the respondents with men making up the remaining 23.2%. This may be consequential given the consistent performance of some of the respondent's above the highest item measures. Also, there's a possibility this difference may result in a failure of invariance.

The Rasch model always reflects the ideal (Linacre, 2012) with the actual obtained results more often than not violating this model. Here, Linacre (2012) emphasizes that while "Many types of violation are inconsequential," there are a few that that have "...serious substantive consequences." So, a possible further analysis using Linacre's WINSTEPS Differential Test Functioning (DTF) functionality to establish whether the TPQ measures function the same way for both genders is required. This involves measuring both genders and then comparing the two sets of difficulties. Should the invariance across these groups not be compromised, it would imply that there is no measure bias due to gender and that the stronger respondents on the item-person maps are due to random reasons. More importantly, it would affirm that some stronger trait items may be needed to maximise the effective range of each of the five measures. This should only be done after the previously proposed adjustments to items 12, 22, 10, and 17 are made and those measures reevaluated.

The results of this study have implications for future behavioural measurement in general, and personality assessment research, specifically.

Traditionally, when constructing *measures* of "the human condition", psychologists and social scientists are unwittingly applying statistical procedures to counts rather than developing measures (Bond and Fox, 2007). This has serious, inexcusable consequences for future people measurement practices. Wright and Linacre (1989) highlight the inadequacies of these traditional approaches by pointing out that 'meaningful measurement is based on the arithmetical properties of interval scales.'

To this end, this research could help the traditional assessment proponents appreciate the importance of the fundamental expectation of Linearity and Conjoint Additivity when constructing future measures and re-evaluating prior practice. That is, to take cognisance of the stark unavoidable fact that, "The construction of measures is a prerequisite of statistical analysis" (Bond & Fox, 2007).

It should also discourage the continued practice of aligning existing assessments with the five factor model or attempting to preserve presented validation procedures by comparing approximated correlations between raw score methodology to Rasch measures. In response to these cautiously pessimistic Rasch antagonists' use of correlations between raw scores and Rasch measures to argue that raw scores are interval, Mike Linacre (2012) writes that, "Treating raw scores as interval measures is like driving in the fog: if the road is straight, you can succeed, but it's slow and difficult. It's easy to drift off the road or take a wrong turn without knowing it, and, if there's anything coming the other way, the outcome can be catastrophic."

In conclusion, the TPQ shows that using Rasch methodology, as a primary

point of departure during the instrument development process, is a necessary and fundamental step for human measurement to be quantifiable and meaningful.

6. Acknowledgments

I would like to thank Mike Linacre for his availability to clarify some of my research results associated with his WINSTEPS 3.92.1 program and all the respondents for taking the time to contribute to the research.

7. References

1. Andrich, D. **A rating formulation for ordered response categories.** Psychometrika, Vol. 43, 1978, pp. 561-574.
2. Bond, T.G. and Fox, C.M. **Applying the Rasch Model: Fundamental Measurement in Human Sciences.** London: Laurence Erlbaum Associates, Publishers, 2001
3. Buchanan, T., Johnson, J.A. and Goldberg, L.R. **Implementing a Five-Factor Personality Inventory for Use on the Internet.** European Journal of Psychological Assessment, Vol. 21, No. 2, 2005, pp. 115-127.
4. Goldberg, L.R., and Saucier, G. **Mapping personality trait structure,** National Institute of Mental Health, U. S. Public Health Service. 1993
5. Goldberg, L. R. **A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models.** In Mervielde, I., Deary, I., De Fruyt, F. and Ostendorf, F. (Eds.), "Personality Psychology in Europe", Vol. 7, pp. 7-28. Tilburg, The Netherlands: Tilburg University Press, 1999
6. Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R. and Gough, H. C. **The International Personality Item Pool and the future of public-domain personality measures.** Journal of Research in Personality, Vol. 40, 2006, pp. 84-96.
7. Linacre, J.M. **Many-faceted Rasch Measurement.** Chicago: MESA Press, 1989
8. Linacre, J.M. and Wright, B.D. **FACETS: Many-Faceted Rasch Analysis.** Chicago: MESA Press, 1997
9. Loevinger, J. **Person and population as psychometric concepts.** Psychological Review, vol. 72, 1965, pp. 143-155.
10. Luce, R.D. and Tukey, J.W. **Simultaneous conjoint measurement.** Journal of Mathematical Psychology, Vol. 1, 1964, pp. 1-27.
11. Rasch, G. **Probabilistic Models for Some Intelligence and Attainment Tests.** Copenhagen: Danish Institute for Educational Research. Chapters V-VII, X, 1960
12. Rasch, G. **On general laws and the meaning of measurement in psychology.** In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics. Berkeley: University of California Press, IV, 1961, pp. 321-334.
13. Rasch, G. **An individualistic approach to item analysis. In Readings in**

mathematical social science. Edited by Lazarsfeld and Henry. Chicago: Science Research Associates Inc., 1966, pp. 89-107

14. Rasch, G. **An item analysis which takes individual differences into account.** British Journal of Mathematical and Statistical Psychology, vol. 19, Part 1, 1966, pp. 49-57.
15. Sitgreaves, R. **Review of probabilistic models for some intelligence and attainment tests.** Psychometrika, Vol. 28, 1963, pp. 219- 220.
16. Townsend, G.C. **So Who Have We Really Been Hiring?** Journal of Corporate Recruiting Leadership, Vol. 3, No. 2, 2007, pp. 9-12.
17. Townsend, G.C. **Townsend Personality Questionnaire Feedback Training Manual.** Available at https:// independent.academia.edu/ Townsend-Gary, 2005
18. Wright, B.D. **Sample-free test calibration and person measurement.** ETS Invitational Conference on Testing Problems. Princeton NJ: Educational Testing Service, 1967, pp. 84-101.
19. Wright, B.D. and Masters, G.N. **Rating Scale Analysis: Rasch Measurement.** Chicago: MESA PRESS, 1982
20. Wright, B.D. **Additivity in psychological measurement.** In Roskam, E. (Ed.) "Measurement and Personality Assessment". Amsterdam: North-Holland, 1985, pp. 101-112.
21. Wright, B.D. **IRT in the 1990's: Which models work best?** In Linacre, J.M. (Ed.) "Rasch Measurement Transactions" Part 2. Chicago: MESA Press, 1996, pp. 196-200.
22. Wright, B.D. **Measuring and counting.** In Linacre, J.M. (Ed.) "Rasch Measurement Transactions" Part 2. Chicago: MESA Press, 1996, p. 371.
23. Wright, B.D. **Comparing Rasch measurement and factor analysis.** Structural Equation Modeling, Vol. 3, No. 1, 1996, pp. 3-24.
24. Wright, B.D. and Linacre, J.M. **Observations are always ordinal: measures, however, must be interval.** Archives of Physical Medicine and Rehabilitation, Vol. 70, 1989, pp. 857-860.
25. Wright, B.D. and Linacre, J.M. **BIGSTEPS: Rasch Computer Program for All Two Facet Problems.** Chicago: MESA Press, 1997
26. Wright, B. D., and Stone, M. H. **Measurement Essentials.** 2nd Edition. Wilmington, Delaware: WIDE RANGE, INC, 1999
27. * * * **International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences** (http://ipip.ori.org/). Internet Web Site.

8. Resources

8.1. Rasch Measurement Models

1. Adams, R. J., Wilson, M. R., and Wang, W. C. **The multidimensional random coefficients multinomial logit model.** Applied Psychological Measurement, Vol. 21, 1997, pp. 1-24.
2. Andrich, D. **A rating formulation for ordered response categories.** Psy-

chometrika, Vol. 43, 1978, pp. 561-574.

3. Andrich, D. **Rasch models for measurement.** Sage university paper series on quantitative measurement in the social sciences. Newberry Park, CA: Sage Publications, 1988
4. Bond, T. G., and Fox, C. M. **Applying the Rasch model: Fundamental measurement in the human sciences.** London: Erlbaum, 2001
5. Fischer, G. H., and Molenaar, I. W. **Rasch models: Foundations, recent developments, and applications.** New York: Springer-Verlag, 1995
6. Linacre, J. M. **Many-facet Rasch measurement.** Chicago: MESA Press, 1989
7. Masters, G. N. **A Rasch model for partial credit scoring.** Psychometrika, vol. 47, 1982, pp. 149-174.
8. Rasch, G. **Probabilistic models for some intelligence and attainment tests,** Copenhagen: Danish Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago Press), 1960
9. Wright, B. D. and Masters, G. N. **Rating scale analysis.** Chicago: MESA Press, 1982
10. Wright, B. D. and Mok, M. **Rasch models overview.** Journal of Applied Measurement, vol. 1, 2000, pp. 83-106.
11. Wright, B. D. and Stone, M. H. **Best test design.** Chicago: MESA Press, 1979

8.2. Rationale for Using Rasch Models

1. Anderson, E. B. **Sufficient statistics in latent trait models.** Psychometrika, vol. 42, 1977, pp. 69-81.
2. Andrich, D. **Distinctions between assumptions and requirements in measurement in the social sciences.** In Keats, J.A., Taft, R., Heath, R.A. and Lovibond, S.H. (Eds.) "Mathematical and Theoretical Systems", North Holland: Elsevier Science Publishers, 1989, pp. 7-16
3. Andrich, D. **Distinctive and incompatible properties of two common classes of IRT models for grade responses.** Applied Psychological Measurement, vol. 19, 1995, pp. 101-119.
4. Andrich, D. **Controversy and the Rasch model: A characteristics of a scientific revolution.** Paper presented at the meeting of the International Conference on Objective Measurement: Focus on Health Care, Chicago, IL, 2001
5. Andrich, D. **Understanding resistance to the data-model relationship in Rasch's paradigm: A reflection for the next generation.** Journal of Applied Measurement, vol. 3, 2002, pp. 325-359.
6. Bond, T. G. and Fox, C. M. **Applying the Rasch model: Fundamental measurement in the human sciences.** London: Erlbaum, 2001
7. Choppin, B. **Lessons for Psychometrics from Thermometry.** International Journal of Educational Research (formerly Evaluation in Education), vol. 9, 1985, pp. 9-12.
8. Fisher, W. P., Jr. **Scale-free measurement revisited.** Rasch Measurement Transactions, vol. 7, 1993, pp. 272-273. www.rasch.org/rmt/rmt71.htm.
9. Fisher, W. P., Jr. **Opportunism, a first step to inevitability?** Rasch Meas-

urement Transactions, vol. 9, 1995, p. 426. www.rasch.org/rmt/rmt92.htm.

10. Fisher, W. P., Jr. **The Rasch alternative.** Rasch Measurement Transactions, vol. 9, 1996, pp. 466-467. www.rasch.org/rmt/rmt94.htm.
11. Linacre, J. M. **The Rasch model cannot be "disproved"!** Rasch Measurement Transactions, vol. 10, 1996, pp. 512-514 www.rasch.org/rmt/rmt103.htm
12. Perline, R., Wright, B. D. and Wainer, H. **The Rasch model as additive conjoint measurement.** Applied Psychological Measurement, vol. 3, 1979, pp. 237-256.
13. Romanoski, J. and Douglas, G. **Test scores, measurement, and the use of analysis of variance: An historical overview.** Journal of Applied Measurement, vol. 3, 2002, pp. 232-242.
14. Smith, R. M. **Applications of Rasch measurement.** Chicago: MESA Press, 1992
15. Wright, B. D. **Sample-Free Test Calibration and Person Measurement.** In Bloom, B. S. (Chair) "Invitational Conference on Testing Problems". Princeton, NJ: Educational Testing Service, 1967, pp. 84-101. Available at www.rasch.org/memo1.htm.
16. Wright, B. D. **Solving measurement problems with the Rasch model.** Journal of Educational Measurement, vol. 14, no. 2, 1977, pp. 97-116. Available at www.rasch.org/memo42.
17. Wright, B. D. and Linacre, J. M. **Observations are always ordinal; measurements, however, must be interval.** Archives of Physical Medicine and Rehabilitation, vol. 70, 1989, pp. 857-860. Available at www.rasch.org/memo44.htm.
18. Wright, B. D. and Masters, G. N. **Rating scale analysis.** Chicago: MESA Press, 1982
19. Wright, B. D. and Stone, M. H. **Best test design.** Chicago: MESA Press, 1979

8.3. Estimation Methodology

1. Fischer, G. H. and Molenaar, I. W. **Rasch models: Foundations, recent developments, and applications.** New York: Springer-Verlag, 1995
2. Linacre, J. M. **Many-facet Rasch measurement.** Chicago: MESA Press, 1989
3. Linacre, J. M. **Estimation methods for Rasch measures.** Journal of Outcome Measurement, vol. 3, 1999, pp. 382-405.
4. Wright, B. D. and Masters, G. N. **Rating scale analysis.** Chicago: MESA Press, 1982
5. Wright, B. D. and Stone, M. H. **Best test design.** Chicago: MESA Press, 1979

8.4. Assessing Dimensionality and Fit

1. Anderson, E. B. **A goodness-of-fit test for the Rasch model.** Psychometrika, vol. 38, 1973, pp. 123-140.
2. Bond, T. G. and Fox, C. M. **Applying the Rasch model: Fundamental measurement in the human sciences.** London: Erlbaum, 2001

3. Engelhard, Jr., G. **Examining rater errors in the assessment of written composition with a Many-Facet Rasch model.** Journal of Educational Measurement, vol. 31, 1994, pp. 93-112.
4. Engelhard, Jr., G. **Clarification to "Examining rater errors in the assessment of written composition with a Many-Facet Rasch model".** Journal of Educational Measurement, vol. 33, 1996, pp. 115-116.
5. Fischer, G. H. and Molenaar, I. W. **Rasch models: Foundations, recent developments, and applications.** New York: Springer-Verlag, 1995
6. Glas, C. A. W. **The derivation of some tests for the Rasch model from the multinomial distribution.** Psychometrika, vol. 53, 1988, pp. 525-546.
7. Kelderman, H. **Loglinear Rasch model tests.** Psychometrika, vol. 49, 1984, pp. 223-245.
8. Linacre, J. M. **Structure in Rasch residuals: Why principal component analysis?** Rasch Measurement Transactions, vol. 12, 1989a, p. 636.
9. Linacre, J. M. **Detecting multidimensionality: Which residual data-types works best?** Journal of Outcome Measurement, vol. 2, 1998b, pp. 266-283.
10. Linacre, J. M. **Prioritizing misfit indicators.** Rasch Measurement Transactions, vol. 9, 1992, pp. 422-423.
11. Linacre, J. M. and Wright, B. D. **Chi-square fit statistics.** Rasch Measurement Transactions, vol. 8, 1992, pp. 360-361.
12. Smith, Jr., E. V. **Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals.** Journal of Applied Measurement, vol. 3, 2002, pp. 205-231.
13. Smith, R. M. **IPARM: Item and Person analysis with the Rasch model.** Chicago: MESA Press, 1991a
14. Smith, R. M. **The distributional properties of Rasch item fit statistics.** Educational and Psychological Measurement, vol. 51, 1991b, pp. 541-565.
15. Smith, R. M. **A comparison of methods for determining dimensionality in Rasch measurement.** Structural Equation Modeling, vol. 3, 1996, pp. 25-40.
16. Smith, R. M. **Polytomous mean square fit statistics.** Rasch Measurement Transactions, vol. 10, 1996, pp. 516-517.
17. Smith, R. M. **Fit analysis in latent trait measurement models.** Journal of Applied Measurement, vol. 1, 2000, pp. 199-218.
18. Smith, R. M., Schumacker, R. E. and Bush, M. J. **Using item mean squares to evaluate fit to the Rasch model.** Journal of Outcome Measurement, vol. 2, 1998, pp. 66-78.
19. Wright, B. D. **Diagnosing misfit.** Rasch Measurement Transactions, vol. 5, 1991a, p.156.
20. Wright, B. D. **Factor item analysis versus Rasch item analysis.** Rasch Measurement Transactions, vol. 5, 1991b, pp. 134-135.
21. Wright, B. D. **Comparing Rasch measurement and factor analysis.** Struc-

tural Equation Modeling, vol. 3, 1996a, pp. 3-24.
22. Wright, B. D. **Local dependence, correlation, and principal components.** Rasch Measurement Transactions, vol. 10, 1996b, pp. 509-511.
23. Wright, B. D. and Linacre, J. M. **Reasonable mean-square fit values.** Rasch Measurement Transactions, vol. 8, 1994, p. 370.
24. Wright, B. D. and Masters, G. N. **Rating scale analysis.** Chicago: MESA Press, 1982
25. Wright, B. D. and Stone, M. H. **Best test design.** Chicago: MESA Press, 1979

8.5. Rating Scale Category Effectiveness

1. Andrich, D. **Category ordering and their utility.** Rasch Measurement Transactions, vol. 9, 1996, pp. 465-466.
2. Andrich, D. **Thresholds, steps, and rating scale conceptualization.** Rasch Measurement Transactions, vol. 12, 1998, pp. 648-649.
3. Linacre, J. M. **Step disordering and Thurstone thresholds.** Rasch Measurement Transactions, vol. 5, 1991, p. 171.
4. Linacre, J. M. **Investigating rating scale category utility.** Journal of Outcome Measurement, vol. 3, 1999, pp. 102-122.
5. Linacre, J. M. **Optimizing rating scale category effectiveness.** Journal of Applied Measurement, vol. 3, 2002, pp. 86-106.
6. Stone, M. and Wright, B. D. **Maximizing rating scale information.** Rasch Measurement Transactions, vol. 8, 1994, p. 386.
7. Wright, B. D. and Linacre, J. M. **Disordered steps?** Rasch Measurement Transactions, vol. 6, 1992, p. 225.
8. Wright, B. D. and Masters, G. N. **Rating scale analysis.** Chicago: MESA Press, 1982
9. Zhu, W., Updyke, W. F. and Lewandowski, C. **Post-hoc Rasch analysis of optimal categorization of an ordered-response scale.** Journal of Outcome Measurement, vol. 1, 1997, pp. 286-304.

8.6. Reliability and Validity

1. Fisher, Jr., W. P. **The Rasch debate: Validity and revolution in educational measurement.** In Wilson, M. (Ed.) "Objective measurement: Theory into practice", Norwood: Ablex Publishing Corporation, Vol. 2, 1994, pp.36-72.
2. Fisher, Jr., W. P. **Is content validity valid?** Rasch Measurement Transactions, vol. 11, 1997, p. 548.
3. Linacre, J. M. **Rasch-based Generalizability theory.** Rasch Measurement Transactions, vol. 7, 1993, pp. 283-284.
4. Linacre, J. M. **Reliability and separation monograms.** Rasch Measurement Transactions, vol. 9, 1995, p. 421.
5. Linacre, J. M. **True-score reliability or Rasch statistical validity?** Rasch Measurement Transactions, vol. 9, 1996, pp. 455-456.
6. Linacre, J. M. **Relating Cronbach and Rasch reliabilities.** Rasch Measurement Transactions, vol. 13, 1999, p. 696.

7. Smith, Jr., E. V. **Reliability of measures and validity of measure interpretation: A Rasch measurement perspective.** Journal of Applied Measurement, vol. 2, 2001, pp. 281-311.
8. Wright, B. D. **Which standard error? Rasch Measurement Transactions,** vol. 9, 1995, pp. 436-437.
9. Wright, B. D. **Reliability and separation.** Rasch Measurement Transactions, vol. 9, 1996, p. 472.
10. Wright, B. D. **Interpreting reliabilities.** Rasch Measurement Transactions, vol. 11, 1998, p. 602.
11. Wright, B. D. and Masters, G. N. **Rating scale analysis.** Chicago: MESA Press, 1982
12. Wright, B. D. and Stone, M. H. **Best test design.** Chicago: MESA Press, 1979

8.7. Metric Development and Score Reporting

1. Linacre, J. M. **Instantaneous measurement and diagnosis.** In Smith, R.M. (Ed.) "Physical Medicine and Rehabilitation State of the Art Reviews", Outcome Measurement Philadelphia: Hanley & Belfus, Inc, vol. 11, 1997, pp.315-324.
2. Ludlow, L. H. and Haley, S. M. **Rasch model logits: Interpretation, use, and transformations.** Educational and Psychological Measurement, vol. 55, 1995, pp. 967-975.
3. Smith, Jr., E. V. **Metric development and score reporting in Rasch measurement.** Journal of Applied Measurement, vol. 1, 2000, pp. 303-326.
4. Smith, R. M. **Applications of Rasch Measurement.** Chicago: MESA Press, 1992
5. Smith, R. M. **IPARM: Item and Person analysis with the Rasch model.** Chicago: MESA Press, 1991
6. Smith, R. M. **Person response maps for rating scales.** Rasch Measurement Transactions, vol. 8, 1994, pp. 372-373.
7. Stanek, J., and Lopez, W. **Explaining variables.** Rasch Measurement Transactions, vol. 10, 1996, pp. 518-519.
8. Woodcock, R. W. **What can Rasch-based score convey about a person's test performance?** In Embretson, S. E. and Hershberger, S. L. (Eds.) "The new rules of measurement: What every psychologist and educator should know". Mahwah, NJ: Erlbaum, 1999
9. Wright, B. D., Mead, R. J. and Ludlow, L. H. **Kidmap: Research memorandum number 29.** Chicago: MESA Press, 1980
10. Wright, B. D. and Stone, M. H. **Best test design.** Chicago: MESA Press, 1979
11. Zhu, W. **Communicating measurement.** Rasch Measurement Transactions, vol. 9, 1995, pp. 437-438.