Reducing the Length of a Goldberg Based Personality Questionnaire Using Item Response Theory & Confirmatory Factor Analysis

A thesis presented in partial fulfilment of the requirements for the degree of

Masters of Science

in

Psychology

at Massey University, Albany,

New Zealand.

Nathan Conrad Phillips

2006

Supervised by

Dr Richard Fletcher

Abstract

Objectives: This study seeks to reconstitute an existing personality questionnaire by identifying the items that capture the best quality information as measured through *Item Response Theory* (IRT). This process will reduce the length of this measure and increase its measurement precision.

Method: A polytomous IRT model (*Graded Response*: Samejima, 1969) will be used to assess the psychometric properties of each item in this questionnaire and produce item level graphs in order to select the best three items for each of the 26 first-order factors. *Confirmatory Factor Analysis* (CFA) will be used to assess the model fit and unidimensionality before and after the IRT selections are made. This will illustrate the improvement gained through both the deletion of redundant items and the selection of high-quality items.

Results: This questionnaire was reduced from 246 items down to 78 items with three high-quality items identified for each of the 26 first-order factors. The model fit considerably improved through this selection process and the reduction of information was minimal in comparison to the amount of items that were deleted.

Conclusions: This study illustrated the power of using IRT for test development. The item selections are not only of benefit for the organisation that supplied the data for this study, but also the original developers as well as any other users of these items as they are freely available via an online source.

Acknowledgements:

I would like to thank the CEO of the New Zealand Organisational Psychology Consultancy who gave me the motivation to begin my thesis, and then the data to make it happen. I am very grateful for your kind words of guidance, as my progression down this path may not have happened without that first meeting.

I would also like to take this opportunity to thank my supervisor, Dr Richard Fletcher, who has assisted me throughout this Masters thesis. His knowledge and direction have guided me towards completing a project of which I am proud to submit.

Dr Linda Jones, from the Wellington Massey University campus, has been of great assistance to me with the important administration side of university life. Dr Jones showed a genuine desire in helping this process run smoothly. Thank you.

Dedication:

I dedicate this thesis to my wife Lara, who has been highly supportive of me throughout this long process.
Lara helped me keep on track and I would not have finished this without her.
This was not an easy topic to get my head around let alone try to explain to friends and family.
Thank you Lara for your love and support at all stages of my life.
I think it is finally time for me to start work!

Table of Contents

List of Tables

List of Figures

List of Appendices

Introduction

Personality Testing

Personality is assessed through determining and measuring individual characteristics or traits that represent important differences between people (Ozer & Reise, 1994). Personality is also viewed as being relatively stable across situations and across time and therefore has many applications if measured in an appropriate manner.

A focal reason for the study of personality stems from the desire to scientifically understand human behaviour. The use of this information is largely of interest to psychologists and other behavioural researchers, but it is also of great interest to organisations. Meta-analyses have illustrated the importance of the relationship between certain personality characteristics and organisational outcomes (Barrick & Mount, 1991; Barrick, Mount, & Judge, 2001). By understanding these relationships organisations are better equipped to seek further information about applicants for positions or promotions and therefore make better decisions.

The Development of Personality Testing

Many questionnaires have been designed and continuously refined to improve the usefulness of their output and the efficiency of their input (Costa & McCrae, 1997). This refinement process began over 60 years ago with the foundation being laid by Raymond B. Cattell (Goldberg, 1990). Cattell was one of the first scientists to apply empirical procedures to the task of constructing a taxonomy of personality items, and achieved this by assessing the correlations amongst the items and by using oblique rotational procedures (Goldberg, 1990).

Cattell (1943) worked to define a short list of categories that encompassed thousands of English personality characteristic adjectives and concluded that the 171 scales he developed could parsimoniously be grouped into a dozen different categories. The academic consensus that followed Cattell's foundation work was that the immense list of items could be grouped under five major headings (Goldberg, 1990).

The Big-Five and Five-Factor Models

The understanding of personality through the measurement of personality traits is widely accepted with the dominant method utilising the five factors alluded to above (Ozer & Reise, 1994). The term applied to this form of grouping is the *Five-Factor Model* (FFM; Guenole & Chernyshenko, 2005) with the most common FFM referred to as the *Big-Five* (Goldberg, 1990). The categories used for tests such as these are traditionally numbered and labelled as follows: (1) *Surgency* or *Extraversion*, (2) *Agreeableness*, (3) *Conscientiousness* or *Dependability*, (4) *Emotional Stability* or *Neuroticism*, and (5) *Culture* or *Intellect* or *Openness* (Goldberg, 1990). These five factors have been shown to account for a large proportion of the variance in self-report personality questionnaires (Guenole & Chernyshenko, 2005) meaning that these fives factors give a good overall impression of an individual's personality. For a full discussion of the history of the *Big-Five* see Goldberg (1990).

Typically, personality questionnaires are lengthy and an excessive amount of time can be spent completing the measure, entering the data, and interpreting the results. Due to the labour involved in this process developers are often requested to reduce the length of questionnaires and by some means maximise the resulting information (Wang, Chen, & Cheng, 2004).

As mentioned by Tuerlinckx, Boeck, and Lens (2002) the accuracy of information provided by lengthy questionnaires comes into question for two main reasons: from the developer's perspective longer questionnaires tend to include lower quality items such as filler items, non-specific items, and items that are included solely to improve reliability; from the participant's perspective longer questionnaires increase the likelihood of losing concentration and making inaccurate responses through boredom, laziness, or unknowingly responding in a repetitive manner. Tuerlinckx et al. also found that questionnaire length significantly correlated with the final score on their measure. They suggested that IRT models could be fitted to personality checklists in a way that could identify a point where test fatigue influences the responses of the participant. They termed this the 'drop-out' point and this was explained as a consequence of loss of attention and loss of patience as participants responded without having fully read the question.

To alleviate the issues that arise out of lengthy questionnaires this research seeks to improve the quality and measurement precision of an existing personality questionnaire by reducing it to the core items that provide the best information about the participant.

The questionnaire that will be used in this research is derived from the freely available online resource at http://www.ipip.ori.org/ipip developed by Goldberg (1990). Many researchers have used this resource (e.g. Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001b; Guenole & Chernyshenko, 2005) including the organisation that provided the data for this study.

The original developers of these items indicated that these are preliminary items as only rudimentary procedures were applied in developing the scales (Goldberg, Johnson, Eber, Hogan, Ashton, Cloninger, & Gough, 2006). Goldberg et al. (2006) suggested that an IRT analysis would identify the highest quality items from this item-pool, and subsequently invited other researchers to perform this task. The results of such an analysis would be applicable for anyone who uses the items from their website however a preliminary search through the 100 plus articles on their website showed no indication of this task being achieved.

Test development has traditionally been performed using *Classical Test Theory* (CTT). However as questionnaires are completed at the item level, it is logical that they should also be developed and interpreted at the item level (Fletcher & Hattie, 2005). This form of analysis cannot be achieved through CTT and therefore an alternative method is necessary.

Theories Underlying Personality Test Development

Classical Test Theory and Item Response Theory

Hambleton and Jones (1993) compare and discuss the two major theories underlying test construction and development, CTT and IRT. They state that models cannot perfectly represent the test data they are associated with, and therefore the question in relation to which theory to use should be based on which will help create a model that will best guide the measurement process. The model strength is dependent on the assumptions that must be met in order to use the relevant framework. Hambleton and

Jones (1993) state that CTT models are often weak as the assumptions are easily met whereas IRT models are stronger as the assumptions are harder to meet. For example, in IRT the assumption is made that the set of items grouped under one label must only measure that single trait or ability and therefore unidimensionality (discussed below) must be satisfied when applying this theory (Raju, Laffitte, & Byrne, 2002). Conversely, CTT only assumes that the structure of a model is consistent when tested with different samples.

The majority of test development is currently performed using CTT. This is due to two main factors. Firstly, IRT is a statistically complex procedure and software was not available that made the process simple to utilise (McKinley, 1989), however this has now changed. Secondly, any new theory must be thoroughly tested and refined before it is applied to real data (Zickar, 1998). IRT has now gone through this process and can thus be used in mainstream testing. IRT has made big impacts on quantitative psychology as the underlying statistical base of IRT along with the development of computer technology has meant that computerised adaptive testing can now be performed. This combination gives the precision of classical tests with the efficiency of advanced software that can select an item that will obtain the most useful information (Ozer & Reise, 1994; Zickar, 1998). The key differences between CTT and IRT will show why IRT is quickly increasing in popularity and use.

Classical Test Theory

In its basic form CTT utilises three core concepts: the observable *test score*, the unobservable *true score* and the unobservable *error score*. CTT provides the assumption that the average *error score* (for the population that completed the test) is zero and hence the *true score* is derived directly from the *test score*. As this assumption is based on the average response to a group of items, two aspects of the data are lost. The first is the ability to assess individual responses, as the output statistics are derived from group averages rather than independent items. The second is that the process of averaging constrains the usefulness of the outcome statistics as no feature of the process indicates that the outcome could be generalised outside the sample from which they were derived, thus making the output statistics 'sample dependent'. As stated by Hambleton and Jones (1993) "this dependency reduces their utility".

Error is not estimated through the CTT procedure. This means that apart from the underlying construct, any other factor that may influence the participant's response is unaccounted for. In contrast, Gefen (2003) explains that every variable in a test introduces an element of measurement error that does not relate to the actual underlying construct. Some CTT models improve upon this basic assumption by indicating that there is measurement error but that the distribution of the error can be estimated using a predetermined curve, such as normal distribution (Hambleton & Jones, 1993). This addition improves the output statistics by identifying the error but does not give a true indication of the error associated with an item.

Item Response Theory

The issues that have been raised in regards to CTT (often produces weak models; loss of item information; sample dependency; unaccounted for error) are overcome by using IRT. IRT is a statistical theory about an individual's response to an item and how that relates to the relevant ability, trait, or construct that is being measured. There are two typical underlying assumptions involved in creating models within the IRT framework. The first pertains to the dimensional structure of the test data (Hambleton & Jones, 1993). This assumes that items that are grouped together are measuring one facet or category of information. This is referred to as *unidimensionality* as each item should only measure one unique factor (McKinley, 1989). The second relates to the form of the graph that represents the item. This graph is created using the data from the item (how people have responded) and a relevant IRT formula (Hambleton & Jones, 1993). The assumption in regards to unidimensionality will now be explained.

Unidimensionality

Gefen (2003) states that every item should only have one underlying construct. This means that items should only reflect their associated construct without significantly reflecting any other. This concept can be clarified through making the distinction between common variance and non-common variance. If two items in a test are hypothesised to measure the same construct then a proportion of the variance they capture is effectively in common. However, items generally do not have perfect measurement properties and therefore also capture other variance that is referred to as non-common variance. An item is not unidimensional when its non-common variance is

highly correlated with the non-common variance of another item, thus indicating that the items are capturing the variance of more than one dimension.

Although this analysis is important for assessing the strength of a model, the literature regarding unidimensionality is controversial. As Hattie (1984, 1985) describes, most indices of unidimensionality have some form of problem. Therefore great care should be taken when selecting which method of analysis is used (for a comprehensive review see Hattie, 1984, 1985). Despite these issues, it is important to assess unidimensionality and this is often performed through *Confirmatory Factor Analysis* (Gefen, 2003).

Item Characteristic Curve

The second IRT assumption pertains to the shape of the graph produced by each individual item. This graph or more specifically the line that is formed by the data on the graph is called the *Item Characteristic Curve.*

The *Item Characteristic Curve* (ICC) is a graphical representation of how and where an item works. The graph plots the probability of a correct response or endorsement of an answer, against ability or endorsement or a trait (McKinley, 1989). The principal of having an ability score is a fundamental difference to CTT that utilises test scores. That is because a person's ability is independent of (1) *the test they are completing*, (2) *the others that complete the test* and (3) *the other items in the test* (Hambleton & Jones, 1993). An example of this is that a person will have a lower score on a difficult test than they will on a simple test, however their ability will remain constant over both tests. Their ability should also remain constant over any other tests that measure the same construct, if completed at the same time. This signifies that ability (or endorsement) can be plotted on a continuum, and this continuum is dependent on the item itself and not the people who responded to the item. This gives all parameters estimated through IRT the property of invariance (McKinley, 1989) and hence the item parameters do not vary when used with different samples.

Sample Independence

McKinley (1989) states, "Item statistics that are obtained from the application of IRT models are independent of the sample of examinees to which a test (or other instrument) is administered". This is in contrast to traditional statistics where scores are stated as a percentage of correct responses and where the statistic most frequently used for comparison is the mean score. This traditional procedure indicates that the output statistics are only relevant to their sample of origin or a sample that has been shown to be very similar. Therefore, in order to obtain comparisons for people completing tests, organisations expend great effort building databases of different sample groups. Conversely, a single analysis can be performed through IRT and all respondents can be assessed on the same scale. In this way IRT avoids sample dependency and adds utility (McKinley, 1989).

IRT Models

There are two major families of IRT models, dichotomous and polytomous. Dichotomous models are for items that have binary answers: *yes* or *no*, *agree* or *disagree*, *1* or *2*. Polytomous models are for items with more than two responses (Ostini & Nering, 2006). Whether dichotomous or polytomous, all IRT models effectively include three estimation parameters: an item discrimination parameter '*a*', a difficulty parameter '*b*', and a guessing parameter '*c*'. In the one-parameter model (or Rasch model) the '*a*' is set at 1, '*c*' is effectively set at 0, and the IRT formula estimates the '*b*' parameter in order to produce the item graphs. In the two-parameter model (or logistic function) both the '*a*' and the '*b*' parameters are estimated by the formula. In the three-parameter model all three parameters are estimated (Baker, 2001; Hambleton, Swaminathan & Rogers, 1991).

IRT gives a true understanding of how an individual item operates through the use of item parameters. The discrimination or '*a*' parameter is labelled as such because it illustrates how well an item differentiates between individuals, as an item with a high '*a*' discriminates more than an item with a low '*a*'. The item difficulty or '*b*' parameter is labelled as such as the item graphs visually illustrate where on the continuum an item operates. Therefore, in regards to ability the value of the '*b*' parameter will indicate whether the item operates in the low end of the scale, hence is an 'easy' item, or the

high end, hence is a 'difficult' item. The '*b*' parameter is also referred to as the 'response option location parameter' (Fletcher & Hattie, 2004) as the graph informs the user where the item best differentiates between individuals, i.e. between people at the low end or high end. It is of benefit to the user to have items in a scale that operate in different areas of the personality continuum. This means more of the information about the latent variable is captured and therefore it can be better understood and is more practical.

An important difference between dichotomous and polytomous models is in regard to the amount of '*b*' parameters that are estimated. In dichotomous models '*b*' represents the threshold point between a respondent choosing category 1 or category 2, e.g. 1 = '*yes*', 2 = '*no*'. However, polytomous models require the estimation of additional '*b*' parameters due to the multiple response options. A polytomous model with, for example, five categories would have four '*b*' parameters labelled '*b1*', '*b2*', '*b3*' and '*b4*', each representing the threshold point between the five category options. A further difference between these models is that the main item graph for dichotomous items is the *Item Characteristic Curve*, whereas for polytomous items this is referred to as the *Category Characteristic Curve* (Fletcher & Hattie, 2004).

There are three main models available when using polytomous IRT. Two of these are *Rasch* type models (one parameter models), namely the *Partial Credit* (PC) model and the *Rating Scale* (RS) model. These only estimate one parameter due to the "Principle of specific objectivity" (Ostini & Nering, 2006), which is derived from the theory that person parameters (which influence the item discrimination parameter) should be separate from the item parameters. Therefore the '*a*' parameter remains constant and only the '*b*' parameters are estimated by the formula (Ostini & Nering, 2006). The PC model assumes that responses are ordered meaning that as a respondent successfully progresses through the items their ability level also increases (Fletcher & Hattie, 2004). The name of this model is due to the fact that a correct response to the first part of an item and not the second part still receives partial credit. The RS model is similar to the PC model and is derived from the same underlying principles (Ostini & Nering, 2006). The third option is the *Graded Response* (GR) model (Samejima, 1969), which does not assume that item discrimination is the same between items.

Theoretically, as the *Rasch* models focus on correct or incorrect responses they are not well suited to personality testing in comparison to the GR model, which is more useful for trait endorsement data. This is illustrated through many studies that have selected this model for the development of personality questionnaires (Bolt, Hare, Vitale, & Newman, 2004; Fletcher & Hattie, 2004; Gomez, Cooper, & Gomez, 2005). For a complete description of the GR model refer to '*Polytomous Item Response Theory Models*' by Ostini and Nering (2006). In addition, a comparison of the application of different IRT models to personality data can be seen in Chernyshenko et al. (2001b).

Typical Methods of Questionnaire Development

Two key aspects of questionnaire development are in regard to (1) *the way in which items fit together in a factor* and (2) *the way factors fit together in a model.* The first of these aspects, item to factor fit, is typically measured through reliability analyses. Churchill (1979) stated that reliability should be the first measure calculated to assess the quality of a factor, the most common measure of which is the *Cronbach's Alpha.* Higher reliability is achieved by having items that load well together. This may signify that the items are asking the same question in a different way. For this reason item to factor fit and also factor to model fit are better measured through unidimensionality analyses such as can be performed through CFA.

CFA has been used for many studies assessing the fit of models for personality inventories (Borkenau & Ostendorf, 1990; Raju et al., 2002; Guenole & Chernyshenko, 2005). In these studies CFA has been stated as an appropriate methodology for confirming the underlying structure of an inventory. An important aspect of these analyses is that they are performed not only to confirm the hypothesised structure, but also to reject other plausible models. Additionally, CFA provides the means to test for unidimensionality, which is of critical importance for test validity (Gefen, 2003). If unidimensionality is not satisfied this can lead to incorrect interpretations of the strength of relationships within the model (Chernyshenko, Stark, & Chan, 2001a). The primary concern addressed through CFA in personality literature is the factor structure of each questionnaire, as there are many opinions regarding which factor structure best describes personality data.

Factor structure disagreement has been a major catalyst for the different forms of personality questionnaires currently available. This conflict is mainly caused by the difference of opinion in regard to what is actually being measured (Eysenck, 1992). Eysenck is the primary personality theorist opposing the FFM and alternatively proposes a three-factor model using *Extraversion, Neuroticism* and *Psychotocism* (Guenole & Chernyshenko, 2005). Ones and Viswesvaran (1998) propose a two-factor model where *Conscientiousness, Agreeableness, Emotional stability* load on one factor, and *Extraversion* and *Intellect* on the second. The 16PF (Conn & Rieke, 1994) is a FFM however the emphasis is on the 16 lower-order factors rather than the five higher-order factors (Chernyshenko et al., 2001a). In each of these cases the factor structure is proposed based on developer preference.

Researchers who question the validity of the design of other measures often test the proposed factor structures with their own data. Chernyshenko et al. (2001a) state that although the 16PF is the most influential and well-researched self-report personality inventory developed in the past 50 years, there was still a need for the unidimensionality of the 16 non-cognitive scales in the 16PF and the hierarchical factor structure of the inventory to be investigated. This was motivated by the recent development of the test from the fourth to the fifth edition as many of the items had considerably changed. Some had minor changes (such as subtle rewording) and many had been discarded and replaced with items that were completely new to the measure. Only 22% of the 185 items in the measure were exactly as they were in the fourth edition, therefore it was of determined that the factor structure should be reconfirmed. Their analysis using a hierarchical *Exploratory Factor Analysis* resulted in a confirmation of the hypothesised factor structure as the 185 items loaded on 16 first-order factors, which loaded on five second-order factors.

Being in its fifth addition the 16PF is an example of a personality questionnaire that is subject to continuous development and improvement (Gerbing & Tuley, 1991). The item level development of this test means that item properties are theoretically constantly being improved with the additional data providing means for the ongoing analysis. Many questionnaires go through the development process (Costa & McCrae, 1997) as this improvement is of empirical benefit to the end users.

Goldberg's Online Inventory

Another example of an ongoing test development process is seen through the constantly updated public domain instrument developed by Goldberg (1990) available at http://www.ipip.ori.org/ipip. Goldberg has made over 2000 items available for researchers, teachers, students, small organisations, or any person who would like to make use of this item bank. Many of the items are based on the major personality inventories that have been mentioned in this study. The items have been correlated with the original scales, redundant items were discarded based on similar wording to other items, reliability analyses were performed and the items have been categorised for those who wish to use them (a full description of this process is available in Goldberg et al., 2006). This has meant that researchers from around the world can use this resource without cost, so they can confirm or reject their personality research hypotheses. As stated, Goldberg has invited any researchers to develop these items using applications such as IRT in order to improve the quality of these scales.

IRT Research

Current personality research has shown some movement towards analysis with IRT. This is a statistically complex procedure (McKinley, 1989) however the detailed information that is provided is invaluable for those who see the importance of measurement precision.

Fletcher and Hattie (2004) applied IRT to a 70-item *Physical Self-Description Questionnaire* (PSDQ) and identified good items, mediocre items that should be reworded, and poor items that should be discarded due to the limited amount of unique information they provided. Through this process Fletcher and Hattie (2004) showed how to minimise the length of the questionnaire by identifying items that captured the best quality information. This item level analysis is only available through IRT.

A further application of IRT is seen through the development of the *Asian Values Scale* through to the *Asian Values Scale- Revised* (Kim & Hong, 2004). In this analysis it was stated that the original 35-item scale was developed using CTT through reliability and validity analyses. The scale was revised using IRT in an attempt to improve the measurement properties of the scale. Their analysis through the use of the *Rasch Model*

resulted in a reduction from their original list down to 25 items and a reduction of response options from the original 7-point Likert-scale down to a 4-point Likert-scale. Hong, Kim and Wolfe (2005) performed a similar IRT analysis with the use of a *European American Values Scale* (EUVS). In this study the EUVS had 18 items, which had been revised from an original list of 180 items. This original list was then subjected to the IRT analysis and 25 items were selected along with the same reduction of 7 response options down to 4 response options for the EUVS-*Revised*. The results of these two studies stemmed from the valuable item level information that was gained through the use of the IRT graphs. It is also interesting to note that in regards to the Likert-scale both of these personality analyses were reduced from 7 options down to 4 options.

Although Kim and Hong (2004) and Hong et al. (2005) opted for a scale wide response option reduction, this is not always the case. Through the IRT analysis performed by Fletcher and Hattie (2004), no changes were made to the questionnaire however recommendations were given. These included items that should be kept as the core of a future revised questionnaire, items that suited the current Likert-scale, items that would be better suited to a dichotomous scale, and items that needed rewording and retesting in order to be included in the revised questionnaire. Fletcher and Hattie (2004) utilised Samejima's GR model, which estimates all three parameters involved in polytomous IRT, whereas Kim and Hong (2004) and Hong et al. (2005) selected the one-parameter Rasch model. Better quality information is typically gained through using the three-parameter model over the one-parameter model, however a larger sample size is needed (Tabachnick & Fidell, 2007) which can limit the model selection.

Gomez et al. (2005) also selected Samejima's GR model for their analysis of two behaviour-based scales. Rather than focus on individual items as was shown through the studies mentioned above, Gomez et al. assessed the information captured by the whole scale. They found that the items were generally good however they only provided information about their latent traits from the moderately low to the moderately high areas of the continuum thus signifying issues for the psychometric properties of the scales. Recommendations were made for additional items to capture information at each end of the continuum.

IRT was also used in a psychometric analysis performed by Tuerlinckx et al. (2002). An interesting component of this analysis was the decision to split their dataset between males and females and use this as a form of cross-validation. From this procedure they were able to illustrate similar findings between the two separate groups and conclude that the findings from one part of their study cross-validated the findings from another part. In regards to any questionnaire development, the process of cross-validation with different samples is highly recommended (Tabachnick & Fidell, 2007).

Objectives and Hypotheses for the Current Study

Longer questionnaires often include redundant items that can decrease the measurement precision of the test (Tuerlinckx et al., 2002). Through IRT the best quality items in a questionnaire can be identified. Therefore a model produced with items selected through IRT should show much better fit than a model that includes redundant items in terms of both unused response options and items that capture little information (Fletcher & Hattie, 2005). IRT assumes unidimensionality and therefore any factors analysed should be assessed using this principle (Raju et al., 2002). For this reason, this study will perform a test of model-fit on the original questionnaire using CFA (*Model 1*), followed by the deletion of poor items as shown through these analyses, after which another CFA (*Model 2*) will be run in order to measure the improved fit of the model. This will be the first stage of analysis and it is hypothesised that the fit of the model will improve.

In order to further identify and select the best items for each factor, IRT analyses will be performed (see Gomez et al., 2005; Fletcher & Hattie, 2004; Kim & Hong, 2004; Hong et al., 2005; and Chernyshenko et al., 2001b). Tabachnick and Fidell (2007) stated that a minimum of three variables should be used to measure a factor. Accordingly, the three best items will be identified for each lower-order factor and these will be combined for a final reconstituted CFA model (*Model 3*). This will be used for comparisons with the previous two models. It is hypothesised that the model fit will once again significantly improve from this procedure.

This reconstituted questionnaire will show far greater measurement precision than its original state with additional efficiency of use due to its reduced length and its lack of redundant items. As information is additive, the deletion of redundant items will lead to

lower overall information, however due to the selection of high-quality and high-information items it is expected that the information reduction will be minimal in comparison to the item reduction. Furthermore, the results of this identification process will be of great value to any users of Goldberg's online resource due to the fact that the parameter estimates are dependent on the individual items and not the sample that was used in this analysis hence the resulting item selections can be freely generalised and are thus highly relevant to many individuals and organisations.

Method

Participants

This study used data provided by an Organisational Psychology consultancy in New Zealand. The sample consisted of 973 adults, 376 of which were female and 597 of which were male. Participants were aged between 16 and 80 (M = 42.40, SD = 8.93). The majority of participants described their ethnic/cultural background as NZ European (n = 774), followed by Other European (n = 102), Maori (n = 80), Asian (n = 22), Pacific Islander (n = 16), and Other Ethnic Group (n = 11).

Measure

Original State of Questionnaire

This personality questionnaire is hypothesised to be a 3-stage higher-order model. The items in the questionnaire were derived from the online resource developed by Goldberg (1990) and are modelled in the design of the *Big-Five*. Therefore, this questionnaire includes five factors that give an indication of personality and these are labelled as follows: *Extraversion and Impact*, *Emotional Management*, *Intellectual Preferences*, *Interpersonal Style*, and *Self Management and Drive*. The relationship between the traditional *Big-Five* factors and the factors in this questionnaire is shown in Fig. 1.

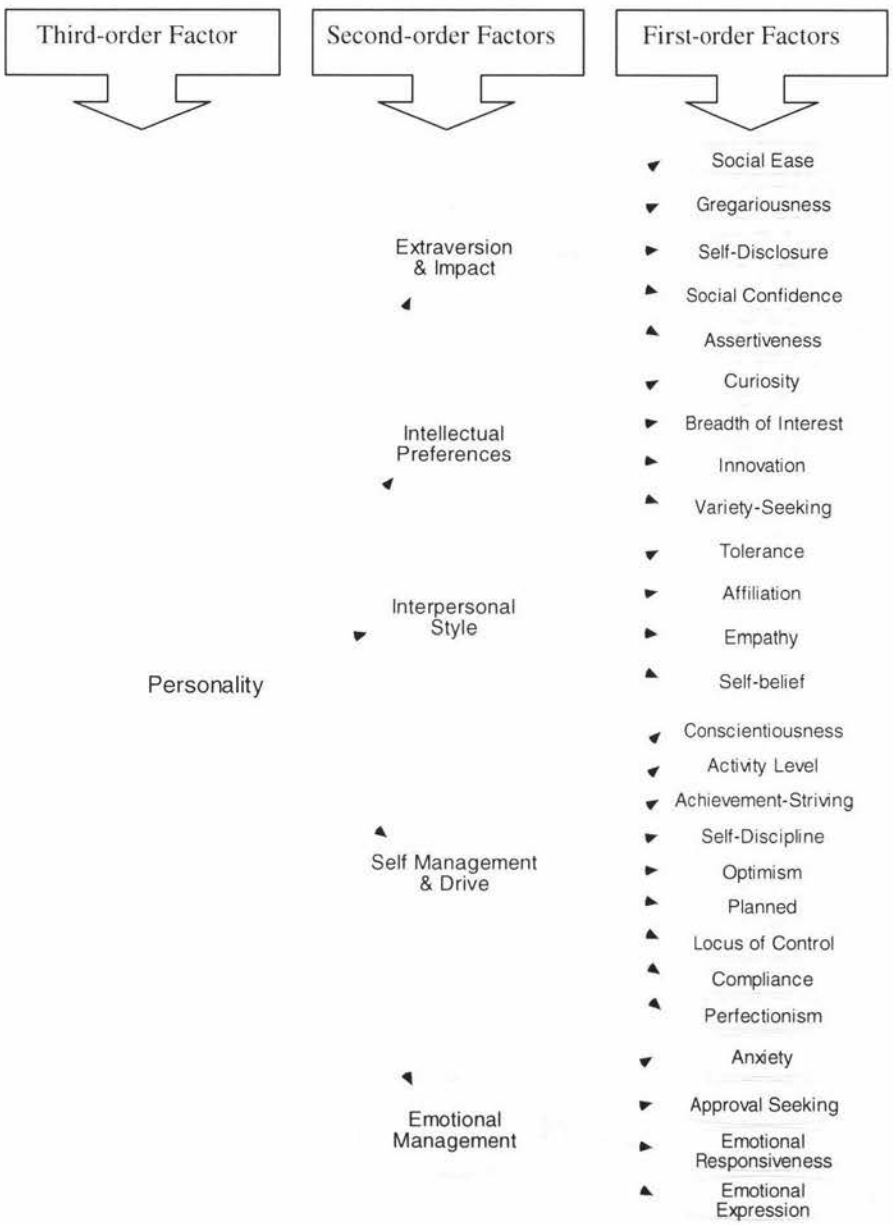
Fig. 1

Relationship between Traditional Big-Five Factors (left) and the Five Second-order Personality Factors (right) in this Questionnaire

A model depicting the hypothesised structure of this personality questionnaire is shown in Fig. 2 with 26 first-order factors, five second-order factors and a single third-order factor. Chernyshenko et al. (2001b) also used this terminology to describe the factors in their analysis of the 16PF.

Fig. 2

Factor Structure of Questionnaire



There were 246 individual items associated with 26 first-order factors in this questionnaire. These were divided between five second-order factors. The amount of items affiliated to each first-order factor is shown in Table 1 and is listed in Appendix 1.

Many of the items in this list cross-loaded to one or more factors, therefore the total number of items shown in Table 1 (268) is greater than the total number of individual items (246). The original questionnaire included a lie scale (17 items) in order to identify participants who did not respond truthfully. As this was not directly associated with the personality factors, it was analysed separately from the rest of the items and is shown at the end of the appendices (Appendix 6).

Table 1

Amount of Items in Each First-order Factor

Second-order Factor	First-order Factor	Items
Extraversion & Impact		
	Social Ease	10
	Gregariousnes	10
	Self-Disclosure	10
	Social-Confidence	10
	Assertiveness	10
Intellectual Preferences		
	Curiosity	10
	Breadth of Interest	10
	Innovation	10
	Variety-Seeking	10
Interpersonal Style		
	Tolerance	11
	Affiliation	10
	Empathy	8
	Self-belief	10
Self-Management & Drive		
	Conscientiousness	10
	Activity Level	10
	Achievement-Striving	10
	Self-Discipline	10
	Optimism	10
	Planned	10
	Locus of Control	20
	Compliance	10
	Perfectionism	9
Emotional Management		
	Anxiety	10
	Approval Seeking	10
	Emotional Responsiveness	10
	Emotional Expression	10
	Total items	268

For all of the items in this questionnaire participants were asked to answer how accurately the item described them using a 5-point Likert-scale:'1'- *Very Inaccurate*, '2'- *Moderately Inaccurate*, '3'- *Neither Inaccurate nor Accurate*, '4'- *Moderately Accurate*, and '5'- *Very Accurate*. As many items were negatively worded (126 out of 268) these were recoded into the same direction as the positively worded items.

Procedure

Split of Dataset

To enhance the validity claims for this measure, the data were randomly split into two files so that the factor structure of the measure could be tested with the full model using the first data set of 484 participants and then with a reduced length scale using the second data set of 489 participants. Cross-validation is a typical procedure used to increase the strength of statistical analyses. For a good example see Tuerlinckx et al. (2002). The full data set of 973 participants was used for the IRT analysis.

Cross-loaded Items

When items represent more than one construct the interpretation of what they represent is difficult to discern. For factor integrity and interpretation items should only load on one factor. This personality questionnaire had 21 items that were suggested to measure more than one factor (Appendix 2). At the beginning of this study a decision was made to discard these items so that the CFA could be run and so that the principle of unidimensionality could be satisfied. The data for these cross-loaded items was not deleted so that they could be reanalysed if any of the first-order factors failed to converge in the initial analysis. Discarding the 21 cross-loaded items left 225 items for *Model 1*. The fit statistics of this model served as a base line for comparisons.

Confirmatory Factor Analysis

To assess the degree of unidimensionality a model was specified for each individual first-order factor resulting in a total of 26 CFAs. Subsequently, the first-order factors were combined with their associated second-order factor in order to create five second-order CFAs. These were then combined with a higher-order *Personality* factor to create the total model that was used for the comparisons (see Fig. 2 above). Three total model CFAs were calculated to illustrate each stage of the development and selection process.

Model 1: The first model is referred to as *Model 1* (225 items) and includes the original length first-order factors after the cross-loaded items were discarded.

Model 2: The *Model 1* first-order factors were then assessed for model fit. Two of the 26 first-order factors failed to converge. The cross-loaded items were added back to

these two scales and they were reanalysed and successfully converged (this process will be explained in the *CFA Results* section). In order to satisfy the requirements of unidimensionality, poor items were deleted from all 26 first-order factors in *Model 1* based on the *Squared Multiple Correlation*. The remaining items from these first-order factors were then reformed into a model referred to as *Model 2* (187 items).

Model 3: The items from *Model 2* were then subjected to the IRT analysis. Three items from each of the 26 first-order factors were selected and combined into a final model referred to as *Model 3* (78 items).

All CFA models were calculated using AMOS 4.0 (Arbuckle, 1999). When an error term was reported in the model to have negative variance the error-variance of the specific parameter was fixed to .001 as is acceptable under these circumstances (Byrne, 2001). Error-variance was fixed to .001 twice for *Model 1*, once for *Model 2*, and twice for *Model 3*.

Model Fit

Fit indices typically reported in *Confirmatory Factor Analyses* are the *Goodness of Fit Index* (GFI: Tanaka & Huba, 1984), the *Tucker-Lewis Index* (TLI: Bollen, 1989) and the *Comparative Fit Index* (CFI: Bentler, 1990), where > .90 indicates adequate model fit for each of these three fit indices. One further fit statistic referred to as one of the best model fit indicators (Fletcher & Hattie, 2004) is the *Root Mean Square Error of Approximation* (RMSEA: Steiger & Lind, 1980), where .00 < .05 indicates close fit, > .05 < .08 indicates reasonable fit, > .08 < .10 indicates tolerable fit, and > .10 indicates poor fit (Browne & Cudeck, 1993). Examination of these fit statistics indicates whether or not a reasonable fit of the data to the model has been achieved.

Item Response Theory

IRT was then used to identify the best three items for each of the 26 first-order factors in *Model 2*. This was achieved using the polytomous GR model (Samejima, 1969). The items selected from the IRT analysis formed *Model 3*.

The Graded Response Model

The GR model was used to produce many different informative graphs and item statistics. These graphs illustrate a wealth of item level information that is not available with traditional statistical analyses (Fletcher & Hattie, 2004). The graphs used in the results section of this study include *Category Characteristic Curves* (CCC), *Operating Characteristic Functions* (OCF), *Item Information Functions* (IIF), and both *Test* and *Scale Information Functions* (TIF & SIF). These graphs illustrate the amount of information captured by each item, first-order factor, second-order factor, and complete model.

Polytomous IRT Graphs

To create the polytomous IRT graphs the '*a*' and '*b*' parameters were extracted from the raw data using a programme developed by Thissen (1991) called MULTILOG 6.0 (BILOG (Mislevy & Bock, 1991) is for dichotomous data; MULTILOG is for polytomous data). Individual files were created in SPSS (Version 13) for each first-order factor using the complete set of data (n = 973). These were converted to files that could be used with MULTILOG so that the discrimination (*a*) and difficulty (*b1*, *b2*, *b3*, and *b4*) parameters for all of the items could be produced (see De Ayala (1993) for a more detailed description of this process). The '*a*' and '*b*' parameters were entered into a MICROSOFT EXCEL spreadsheet and Samejima's (1969) GR model formula was used to produce the graphs.

Method for Selection of Three Best Items for each First-order Factor

Items were selected based on the item properties illustrated in the item level graphs. These properties include:

(1) *the shape of the graph,*

(2) *the location of item information,*

(3) *total item information,*

(4) *the use of all the response options,* and

(5) *the combination of the items in the first-order factor,* including the

   a. *item information location* and the

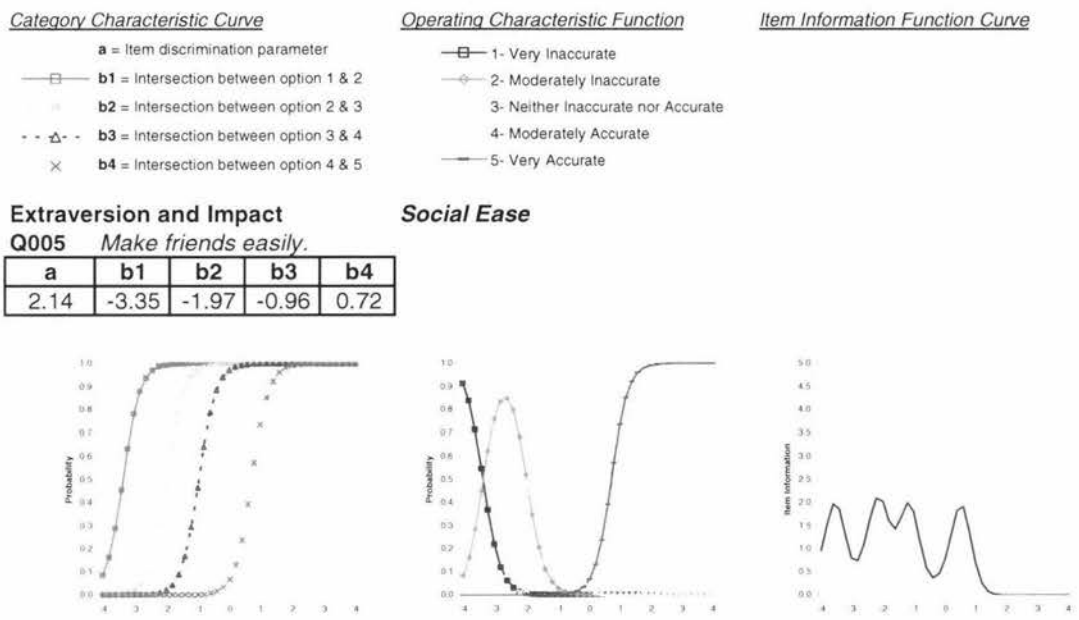
   b. *item wording*

(1) The Shape of the Graph

The *Operating Characteristic Function* (OCF) for item Q005 (Fig. 3) illustrates certain properties that made this a good item. The area under each individual curve is effectively the information that is captured by that response option and in the OCF for item Q005 each individual response option had a high peak that was separate from the other peaks, meaning that each option captured a significant amount of unique information.

Fig. 3

CCC, OCF, and IIF for Social Ease Item Q005



(2) Location of Item Information

Item information should be evenly dispersed between response options and should only operate in a small section of the personality continuum. This means an item can have a more accurate degree of differentiation between individuals on the trait being measured. As seen in the CCC for item Q005 (Fig. 3), this is an example of a good item whereas item Q142 (Fig. 4) is an example of a poor item.

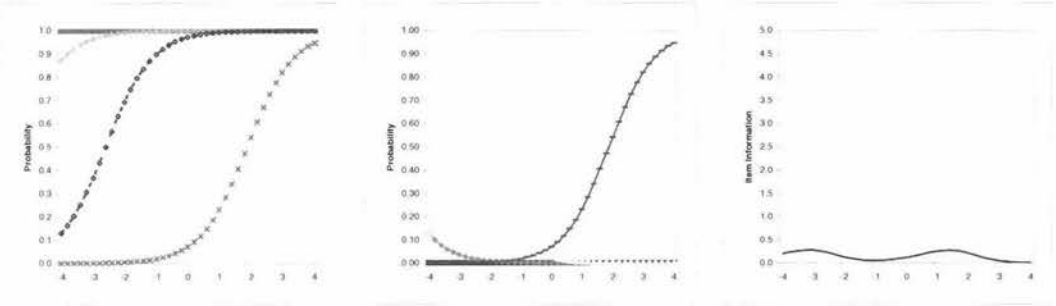
Fig. 4

CCC, OCF, and IIF for Tolerance Item Q142

Interpersonal Style ***Tolerance***

Q142 *Believe that others have good intentions.*

a	b1	b2	b3	b4
0.80	-12.39	-5.41	-2.60	1.88



(3) Total Item Information

An additional aspect to notice when comparing items Q005 and Q142 is that the total information shown in the IIF was considerably lower in item Q142 from the '*Tolerance*' first-order factor. This also indicates that it is a poor item.

Two other items from the '*Tolerance*' first-order factor are shown in Fig. 5.

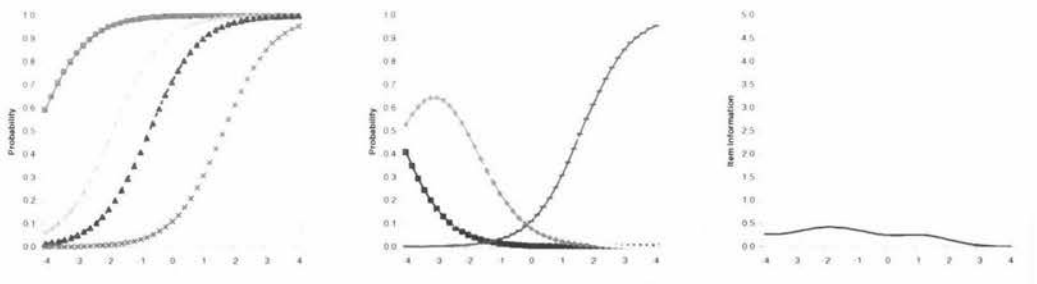
Fig. 5

CCC's, OCF's, and IIF's for Three of the First-order Factor Items for Tolerance

Interpersonal Style ***Tolerance***

Q104 *Am a bad loser.*

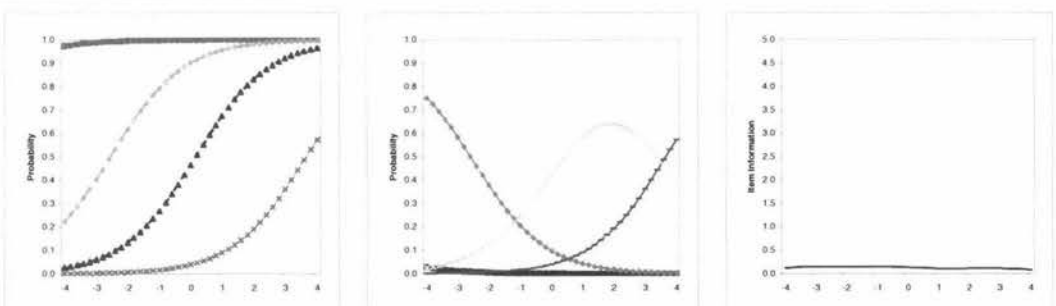
a	b1	b2	b3	b4
0.75	-4.29	-1.88	-0.71	1.64



Interpersonal Style ***Tolerance***

Q164 *Lay down the law to others.*

a	b1	b2	b3	b4
0.51	-8.14	-2.55	0.16	3.66



Bolt et al. (2004) suggest that '*a*' parameters need to be over 1.00 to indicate reasonable discrimination whereas these two items each had '*a*' parameters of 0.80 or less and hence captured very little information. Items that were shown to have information levels similar or worse than these were categorised as poor items.

(4) The Use of All the Response Options

Although Q094 (Fig. 6) has high peaks, high information, and good information location, the OCF shows that the information captured by response option '3' (*Neither Accurate nor Inaccurate*) was also captured by options '2' and '4'. This means that option '3' was effectively redundant and therefore this item was not suited to a 5-point Likert-scale questionnaire. Consequently, items such as this were not selected.

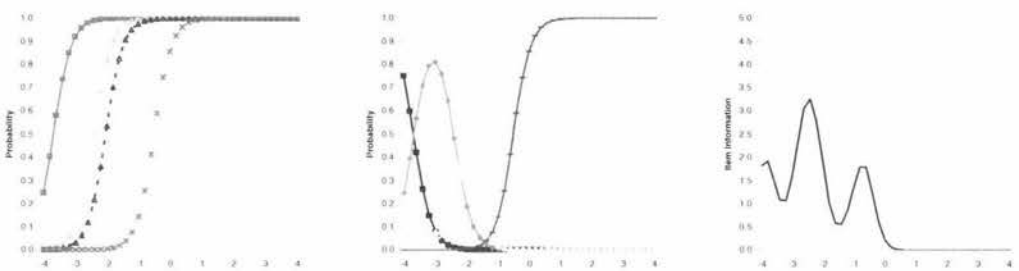
Fig. 6

CCC, OCF, and IIF for Innovation Item Q094

Intellectual Preferences ***Innovation***

Q094 *Can't come up with new ideas.*

a	b1	b2	b3	b4
2.09	-3.69	-2.41	-2.04	-0.5



(5) The Combinations of Items in a First-order Factor

The item combinations were also analysed in order to maximise the variation amongst the wording of the items in each first-order factor and to capture information from different parts of the personality continuum. Examples are given below.

(5a) Information Location: Of the three items seen below in Fig. 7, two were selected for the final model. Although items Q021 and Q026 had higher '*a*' parameters than Q032, the areas under the graphs of Q021 and Q026 (as can be seen in each IIF) were very similar, illustrating that they captured almost the same information. For this reason it was preferred to select only one of these items and then select a different item that captured different information, such as item Q032.

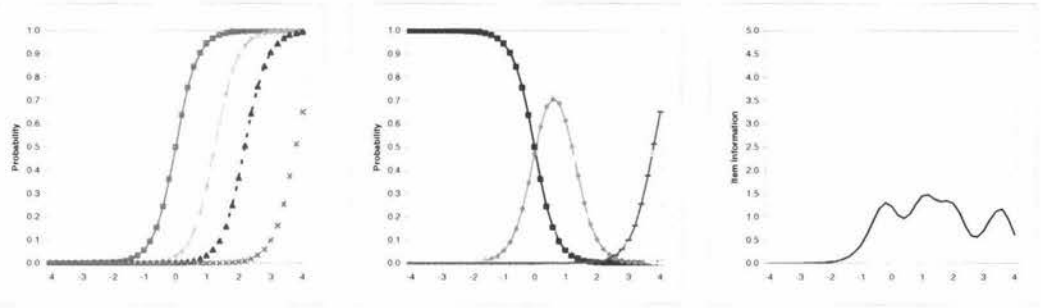
Fig. 7

CCC's, OCF's, and IIF's for Three of the Anxiety First-order Factors Items

Emotional Management ***Anxiety***

Q021 *Have frequent mood swings.*

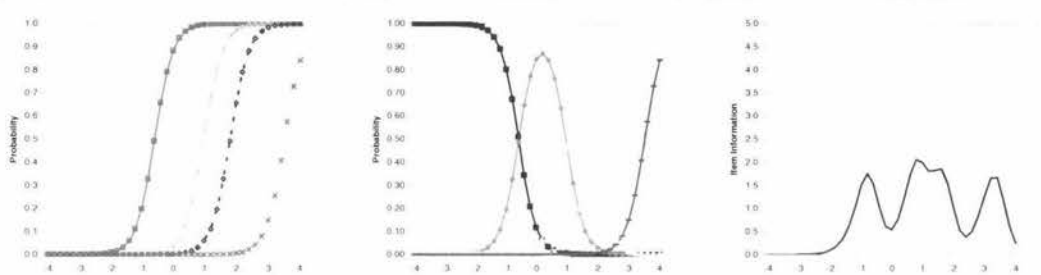
a	b1	b2	b3	b4
1.66	0	1.25	2.19	3.78



Emotional Management ***Anxiety***

Q026 *Get upset easily.*

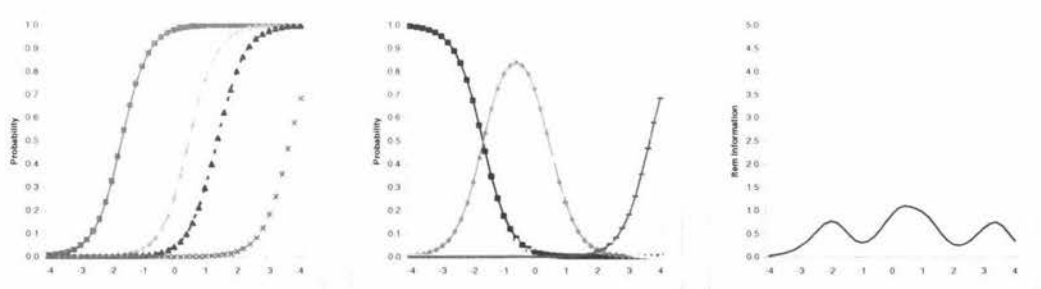
a	b1	b2	b3	b4
2.01	-0.59	0.97	1.81	3.51



Emotional Management ***Anxiety***

Q032 *Am not easily bothered by things.*

a	b1	b2	b3	b4
1.33	-1.68	0.48	1.38	3.66



(5b) Item Wording: The wording of an item also provides insight in regards to which items to select. For the first-order factor '*Empathy*', two of the items were worded as follows: Q116- '*Make people feel welcome*' and Q236- '*Take time out for others*'. The third choice was between two options: Q280- 'Am concerned about others' and Q099- '*Reassure others*'. The highest '*a*' parameter belonged to Q099, however as this item referred to a behaviour, as did the first two, it was rejected in favour of Q280 which refers to an emotion and hence was semantically different.