Influence of Context on Item Parameters in Forced-Choice Personality Assessments

Yin Lin

University of Kent, CEB

Anna Brown

University of Kent

Author Note

This work was supported by the Economic and Social Research Council and CEB (ESRC CASE studentship, grant reference ES/J500148/1).

Correspondence concerning this article should be addressed to Yin Lin, The Pavilion, 1 Atwell Place, Thames Ditton, Surrey, KT7 0NE, United Kingdom. E-mail: yl263@kent.ac.uk

Abstract

A fundamental assumption in computerized adaptive testing (CAT) is that item parameters are invariant with respect to context – items surrounding the administered item. This assumption, however, may not hold in forced-choice (FC) assessments, where explicit comparisons are made between items included in the same block. We empirically examined the influence of context on item parameters by comparing parameter estimates from two FC instruments. The first instrument was compiled of blocks of three items, whereas in the second, the context was manipulated by adding one item to each block, resulting in blocks of four. The item parameter estimates were highly similar. However, a small number of significant deviations were observed, confirming the importance of context when designing adaptive FC assessments. Two patterns of such deviations were identified, and methods to reduce their occurrences in a FC CAT setting were proposed. It was shown that with a small proportion of violations of the parameter invariance assumption, score estimation remained stable.

Keywords: Forced choice, computerized adaptive testing, multidimensional item response theory

Research has related personality to many important life outcomes, ranging from happiness and wellbeing to occupational performance and satisfaction (Ozer & Benet-Martinez, 2006; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). As a result of their utility, personality assessments are popular in many real-life applications (e.g., Barrick & Mount, 1991). Due to the lack of better alternatives in many practical settings, the measurement of personality is almost always conducted using a self-report questionnaire. Traditional personality questionnaires typically share two features: first, single-stimulus (SS) response formats are adopted, asking respondents to describe themselves in relation to a series of items, one at a time, using pre-defined response options; second, multiple traits representing different facets of personality are assessed, each measured by a small, static set of items. Despite the prevalence of these features in traditional personality assessments, the trend is gradually shifting. On one hand, to address some long-standing practical challenges associated with the SS response format, researchers have turned their attention towards an alternative, forced-choice (FC) response format, where respondents describe themselves in relation to a series of items by ranking a number of them at a time instead of rating each individually. On the other hand, thanks to the rise of computer-based testing, there is increasing interest in computerized adaptive testing (CAT) methodologies to improve measurement efficiency – a development that is especially relevant for long, multi-trait personality assessments. The combination of the FC and CAT methodologies, however, is still relatively under-researched. For example, we do not know whether basic assumptions made in the CAT assessments will hold when the FC formats are employed. This study examines one of the essential assumptions for CAT – the assumption of item parameter invariance regardless of the place in a test where that item appears.

Two Challenges of Personality Measurement, and Forced Choice

Self-report personality assessments typically assess multiple traits with a SS response format. For example, the main NEO Personality Inventories, including NEO-PI-R and NEO-PI-3, assess 30 facets of five personality dimensions using 240 items in conjunction with five response options ranging from “strongly disagree” to “strongly agree” (Costa & McCrae, 1992; McCrae, Costa, & Martin, 2005).

A first challenge faced by such traditional personality assessments is widely acknowledged and well documented – its susceptibility to various response biases and distortions. Response biases and distortions arise due to differences in interpretation of the rating scale (Friedman & Amoo, 1999); individual response styles such as central/extreme tendency, acquiescence, socially desirable responding (Paulhus, 1991; Paulhus & Vazire, 2007), and faking (e.g., Donovan, Dwight, & Hurtz, 2003; Griffith, Chmielowski, & Yoshita, 2007). Forcing choice between personality items has emerged as an approach to prevent biases and distortions (Nederhof, 1985; Zavala, 1965). Questionnaires using the FC format place items into blocks and ask respondents to rank the items within the block according to the extent they describe their personality. For example:

Please select one statement that is most true or typical of you, and another statement that is least like you:	Most	Least
I am lively in conversation I persevere with tasks I avoid taking criticism personally		

Research on the FC format has demonstrated that it removes all uniform response biases including central/extreme tendency and acquiescence (Cheung & Chan, 2002) and also provides greater resistance to motivated distortions (e.g., Christiansen, Burns, & Montgomery,

2005). Moreover, the decades-long challenge of overcoming ipsative scores and inferring proper measurement from FC data (Cornwell & Dunlap, 1994; Hicks, 1970; Johnson, Wood, & Blinkhorn, 1988) has been resolved by Item Response Theory (IRT) modeling of FC responses (Brown, 2014; Brown & Maydeu-Olivares, 2011, 2013; Chernyshenko et al., 2009; Stark, Chernyshenko, & Drasgow, 2005). These features of FC thus made it an attractive option for improving assessment fairness and accuracy when biases and distortions are of concern, for example in cross-cultural studies affected by culturally-specific response styles (van de Vijver & Leung, 1997; van Herk, Poortinga, & Verhallen, 2004), and in high-stake assessments affected by faking (e.g., Viswesvaran & Ones, 1999).

A second challenge faced by traditional personality assessments arises from the multidimensional nature of personality, requiring long assessment times to measure all traits reliably. Long assessment times may lead to boredom or fatigue, thus negatively affecting the assessment experience. Modern IRT methodologies address this challenge in two ways. First, by extracting response information in a more efficient manner, assessment length can be shortened significantly. For example, Brown and Bartram (2009) refined a classically-scored FC personality assessment using IRT methodologies, successfully reducing assessment time by 40-50% while maintaining similar levels of score reliability. Second, when combined with computer-based testing technology, IRT opens up the possibility of tailoring the assessment to each and every individual. CAT has demonstrated great utility in the field of cognitive ability testing, with studies showing that tests can be shortened by 50% compared to static paper-and-pencil tests (Embretson & Reise, 2000). Personality assessments nowadays have not integrated CAT methodology as widely or deeply as modern cognitive ability tests, but the prospect is promising. For example, Hol, Vorst and Mellenbergh (2008) conducted a real-data simulation study using SS personality items, and found that CAT only requires as few as 33% of the original items to reach similar levels of measurement accuracy.

A natural next step to advance personality assessments further may thus be to combine FC and CAT methodologies, ensuring good resistance to biases and distortions while also improving measurement efficiency. In a series of simulation studies conducted by Stark and Chernyshenko (2007, 2011) and Stark, Chernyshenko, Drasgow and White (2012), adaptive FC personality tests outperform their static or random counterparts by a large margin, typically reaching the same level of true score correlation at about half the test length. Independently, Brown (2012) conducted a simulation study comparing adaptive and randomized personality tests using a different item bank and a different IRT model, and replicated similar levels of efficiency gain. Some assessments adopting the FC CAT methodology are already in operation, including the Navy Computerized Adaptive Personality Scales (NCAPS; Houston, Borman, Farmer, & Bearden, 2006), the Tailored Adaptive Personality Assessment System (TAPAS; Drasgow et al., 2012), and the Global Personality Inventory-Adaptive (GPI-A; CEB, 2009-2014).

FC CAT and Item Parameter Contextual Invariance

Based on current evidence and trends, it is reasonable to expect that FC CAT methodology will continue to flourish. Nevertheless, the marriage of FC and CAT still has some open questions. In the most unconstrained form of FC CAT, items are adaptively assembled into FC blocks, and the properties of FC blocks are derived from the properties of the constituting items. This simple process requires an item to function in exactly the same way regardless of what other items appear in the same FC block. In IRT terms, this is equivalent to making the assumption that the item parameters are invariant with respect to context – the items surrounding the target item in the FC block.

However, the way items are combined into FC blocks can potentially introduce contextual changes, leading to respondents viewing the items in a different light. The impact

of context on item functioning is neither new nor unique to forced choice. For example, Strack, Martin and Schwarz (1988) showed that by simply swapping the order of two satisfaction items, their correlational relationship changed, producing the item-order effect. At the same period, Knowles (1988) demonstrated that the constructs being measured by a personality assessment become clearer to the respondents as they consider more items, leading to more "polarized, consistent, and reliable" responses in items appearing later in the assessment, producing the serial-order effect. More recently, Steinberg (2001) showed that presenting two SS items on anger experience and anger expression next to each other lead to more extreme responses than when they were presented on their own. Phenomenon as such can lead to change in item properties and thus item parameter shifts.

While item parameter shifts due to change in context are relevant for both linear and CAT assessments, in practice this problem can be fully addressed for linear FC assessments. With a fixed FC form, estimation of item parameters can be done using this particular linear form. In this case, the context (i.e., surrounding items in the same block) remains constant between calibration and application of the assessment. In the more complex case when multiple, parallel linear FC forms with overlapping items are employed, the forms can be calibrated independently and subsequently equated at the form level, without necessarily imposing the parameter invariance assumption on the common items. It is only when the items move blocks from one form to the next, for example in FC CAT or any non-adaptive but dynamic FC assessments, that context differences between calibration and application become inevitable, and thus item parameter invariance becomes a paramount assumption. In other words, there is no guarantee that people will interpret each and every item in a consistent way (leading to invariant item parameters), when other items around it change as in the case of FC CAT.

Empirical studies are needed to examine the effect of context on item functioning. While recent findings (Lin, Inceoglu, & Bartram, 2013) have provided some reassurance on the stability of person parameter estimation when FC block compositions vary, examination of the item parameter stability assumption is still lacking. So far, the justification of this pre-requisite assumption of FC CAT had been ignored by most FC CAT researchers. The research question is whether varying contexts have negligible impact on people's FC responding behaviors and thus on the subsequently deduced item parameters. More specifically, research should quantify the level of item parameter stability when the context around one item is altered due to the presence of other items.

The current study explored the effect of context on item functioning in forced-choice blocks, by examining empirically estimated item parameters across two instruments. The first instruments was compiled of FC blocks of three items, whereas in the second, the context was manipulated by adding one item to each block, resulting in FC blocks of four. The robustness of the parameter invariance assumption required for CAT was examined, and situations where this assumption was violated were identified. Practical strategies to avoid such violations were suggested to inform future forced-choice CAT designs.

Method

Instruments

The Occupational Personality Questionnaire (OPQ32) is an assessment of people's behavioral preference or style in the workplace, providing measurement for 32 traits (Bartram, Brown, Fleck, Inceoglu, & Ward, 2006). The present study utilizes two versions of this assessment that employ a multidimensional forced-choice (MFC) format (i.e., a FC format where items in the same block indicate different traits): the OPQ32i and the OPQ32r. Both versions request respondents to choose the statement that is "most" and "least" like them

within each of the 104 forced-choice blocks. However, the OPQ32i blocks consist of four items (so-called "quads") and OPQ32r blocks consist of three items ("triplets"). The OPQ32r triplets were developed through removing one item per quad from OPQ32i (Brown & Bartram, 2009-2011). Except wording improvements for 5 items, all other remaining items were exactly the same across versions. This nested design allows studying the effect on responding behavior of contextual change caused by an additional, distractor item in the same forced-choice block.

Samples

Historical anonymous data from live administrations of the OPQ32 in UK English in the United Kingdom was used in this study. The samples were collected through a large number of assessment projects, which were typically for employee selection or development purposes. Respondents in the first sample (N=62,639) completed the older, quad instrument between 2004 and 2009. Respondents in the second sample (N=22,610) completed the newer, triplet instrument between 2009 and 2011. As shown in Table 1, the two samples had very similar gender compositions – each had just over 60% males and just under 40% females. In each sample, all working ages were represented, and the majority of respondents were white.

INSERT TABLE 1 NEAR HERE

Analysis Strategy

Analysis was structured in four main steps. Firstly, to create the foundation for all subsequent analyses, item parameters for the quad and triplet instruments were estimated independently using their respective samples, and equated to the same scales in order to remove sample-specific metric differences in the resulting model parameters. Secondly, to examine the impact of instrument design change on people's responding behavior at item

level, item parameters for the quad and triplet instruments were compared directly. Thirdly, to identify underlying reasons of item parameter differences, qualitative contextual analysis of item content was conducted. Finally, to examine the robustness of measurement at trait level, trait score estimates based on different item parameter sets were compared.

Item parameter estimation. The two samples were analyzed using appropriate Item Response Theory (IRT) models. Firstly, the "most" and "least" responses to forced-choice blocks were converted to binary outcomes associated with pairwise comparisons within blocks, as described in Brown and Maydeu-Olivares (2012):

Format	Items	Binary Outcomes
Quad	A, B, C, D	{A,B}, {A,C}, {A,D}, {B,C}, {B,D}, {C,D}
Triplet	A, B, C	{A,B}, {A,C}, {B,C}

The binary outcome of each pairwise comparison $\{i, k\}$ was dummy coded:

$$y_{\{i,k\}} = \begin{cases} 1 & if\ item\ i\ is\ preferred\ over\ item\ k \\ 0 & if\ item\ k\ is\ preferred\ over\ item\ i \\ missing & if\ the\ outcome\ of\ comparison\ is\ unknown \end{cases} \quad (1)$$

Each block of 4 items was coded as 6 pairwise comparisons. The quad instrument thus had 104*6 = 624 binary outcomes. Each block of 3 items was coded as 3 pairwise comparisons, and the triplet instrument had 104*3 = 312 binary outcomes.

Secondly, a Thurstonian IRT model (Brown & Maydeu-Olivares, 2011) with 32 correlated latent traits indicated by their respective observed binary outcomes was fitted to each sample independently using the Unweighted Least Squares estimator in Mplus software (Muthén & Muthén, 1998-2012). The conditional probability for a positive outcome of pairwise comparison was modelled as follows (Brown & Maydeu-Olivares, 2011):

$$Pr\left(y_{\{i,k\}} = 1|\eta_a, \eta_b\right) = \Phi\left(\frac{-\gamma_{\{i,k\}} + \lambda_i\eta_a - \lambda_k\eta_b}{\sqrt{\psi_i^2 + \psi_k^2}}\right). \quad (2)$$

In this expression,

- $y_{\{i,k\}}$ denotes the dummy coded binary outcome for pairwise comparison $\{i, k\}$;
- η_a and η_b denote the latent traits indicated by items *i* and *k* respectively;
- $\gamma_{\{i,k\}}$ denotes the threshold for pairwise comparison $\{i, k\}$;
- λ_i and λ_k denote the factor loading of items *i* and *k* on their respective latent traits;
- ψ_i^2 and ψ_k^2 denote the unique variances of items *i* and *k* respectively.

To enhance parallelism in the comparison of model parameters later, the models only considered binary outcomes shared by both instruments – that is, the 312 binary outcomes as in the triplet version. The outcomes unique to the quad instrument were ignored for two reasons. First, they were not relevant for answering the question of how people's responding behavior changed when a fourth item was added into the same block. The fourth item acted merely as a distractor (context) in the present study's design. It existed only in the quad version, and therefore the parameters relevant to this distractor item could not be estimated for the triplet version, and therefore provided no basis for any parameter comparison. Second, the inclusion of the additional outcome variables when estimating the model parameters for the quad instrument would make the two models non-equivalent, thus introducing an extra source of difference into the comparison of model parameters. The only type of difference of interest to this study was the differences caused by empirical behavior change between the two versions.

The OPQ32 instruments employed a well-established model of workplace personality (Bartram, Brown, Fleck, Inceoglu, & Ward, 2006). Many studies had replicated OPQ32 scale correlations, and found them to be very stable across contexts and even language versions

(for example, see CEB, 2014, Table 15). For the present study, both samples were collected from the same country (United Kingdom), in the original English language version. The IRT scoring protocol applied to UK English OPQ32 data in operational settings uses Bayesian maximum-a-posteriori estimation, informed by the prior distribution of the 32 traits with the correlation matrix established on "a representative sample of the British population collected by the Office of National Statistics in parallel to their Labour Force Survey", and contained 2028 individuals (Bartram et al., 2006; see Table 1). Therefore, the trait correlations in our models were fixed to these same correlations in order to define the factorial space. Furthermore, the origin and unit for each latent trait was set so that the sample's latent trait mean was 0 and standard deviation was 1. For model identification, the unique variance of one item per forced-choice block was fixed arbitrarily to 0.5 (see Brown & Maydeu-Olivares, 2012). To ensure comparability of parameter estimates across instruments, for each corresponding forced-choice block in quad and triplet instruments, the same item was chosen for fixing the unique variance.

However, the partial ranking design of the quad instrument resulted in some missing outcomes that needed additional treatment before item parameters could be estimated. Missing data arose because the "most" and "least" response format did not provide full rank ordering information for blocks of four items – the rank order of the two unselected items was not collected by design. The mechanism was missing at random (MAR), but not missing completely at random (MCAR), since the pattern of missingness was fully determined by the observed responses (Brown & Maydeu-Olivares, 2012). Thurstonian IRT models use limited information estimators (i.e. ULS) based on tetrachoric correlations of the observed binary dummy variables. Previous research by Asparouhov and Muthén (2010) showed that limited information estimators such as the ULS used in the present study result in biased parameter estimates when data were missing at random (MAR) but not completely at random (MCAR).

Because the focus of the present study is on the item parameters, any systematic parameter estimation bias is unacceptable. However, the above bias can be eliminated almost completely using multiple imputation with as few as five replications (Asparouhov & Muthén, 2010). Following the guidance developed specifically for FC data by Brown and Maydeu-Olivares (2012), multiple imputation with 10 replications was applied to handle the MAR data in the quad instrument, in order to prevent any bias in parameter estimation.

Due to the very large size of the quad instrument (416 items, resulting in 624 dummy observed variables), it was not possible to run multiple imputation on the entire instrument all at once. Instead, the quad instrument was divided into 12 similarly-sized subsections covering all 104 forced-choice blocks. Multiple imputation was then conducted using all available data for each of the subsections. Even with this sub-sectioning, due to very large samples used in this study, Bayesian estimation of the unrestricted model required for multiple imputation for each subsection still took up to one day to complete. A total of 10 samples were imputed for each subsection and the resulting data subsequently merged across subsections to reconstruct the complete instrument. Thurstonian IRT model was then fitted to each of the 10 imputed samples. All 10 models converged and gave expected parameter estimates, which were stable across imputations (see the imputation statistics in Table 2). The estimates from the 10 models were then averaged to give the final IRT parameter estimates for the quad instrument.

INSERT TABLE 2 NEAR HERE

Item parameter equating. The parameters of a multidimensional IRT model have a degree of arbitrariness – they are indeterminant until the trait directions, origins and units have been fixed (Reckase, 2009, p. 233-234). In the present study, the IRT models for the triplet and quad instruments were constructed using two different samples. To identify trait directions, both models were estimated while fixing the correlations between latent traits,

thus ensuring identical factorial space. To identify latent trait metrics for each model, we fixed the latent trait origins and units to reflect the means and standard deviations of the individual samples. However, the two samples were far from randomly-equivalent, and thus we fully expected the resulting latent trait metrics of the two models to be different. As a result, the item parameters of the two models were not directly comparable. Therefore, equating was required to place the item parameters on the same scale before subsequent analyses and comparisons.

The Thurstonian IRT model describing responses to forced-choice questionnaires is a variant of the multidimensional 2-parameter normal-ogive (M2PNO) model with some special features. The special features include some constraints on the factor loadings (factor loadings for pairs involving the same item are equal), and correlated error structures for pairs involving the same items. The latter feature allows separate identification of unique variance parameters (in M2PNO model, error variances are all fixed to 1 for model identification). Metric transformation equations for the M2PNO model have long been published (e.g., Davey, Oshima & Lee, 1996). For the Thurstonian IRT model, however, additional attention is needed to handle the unique variance parameters, thus demanding the deduction of new metric transformation equations. We provide these in the present paper.

With latent trait directions fixed to be equivalent across models, transforming of origins and units could be captured by a linear transformation as per unidimensional equating (Kolen & Brennan, 2004, p. 162):

$$\eta_a^* = x_a \eta_a + y_a. \tag{3}$$

In the present study, the aim of equating was to find optimal coefficients x_a and y_a to transform the metric of the quad instrument model (η_a) to the metric of the triplet instrument model (η_a^*).

Transforming the metric of latent traits has implications on item parameter values. For the IRT model to be invariant after transformation, the conditional probability of responses needs to remain unchanged (Reckase, 2009, p. 235):

$$Pr\left(y_{\{i,k\}} = 1|\eta_a, \eta_b\right) = \Phi\left(\frac{-\gamma_{\{i,k\}} + \lambda_i\eta_a - \lambda_k\eta_b}{\sqrt{\psi_i^2 + \psi_k^2}}\right) = \Phi\left(\frac{-\gamma^*_{\{i.k\}} + \lambda^*_i\eta^*_a - \lambda^*_k\eta^*_b}{\sqrt{\psi_i^{*2} + \psi_k^{*2}}}\right) =$$

$$= \Phi\left(\frac{-\gamma^*_{\{i,k\}} + \lambda^*_i(x_a\eta_a + y_a) - \lambda^*_k(x_b\eta_b + y_b)}{\sqrt{\psi_i^{*2} + \psi_k^{*2}}}\right) =$$

$$= \Phi\left(\frac{-\gamma^*_{\{i,k\}} + \lambda^*_i y_a - \lambda^*_k y_b + \lambda^*_i x_a\eta_a - \lambda^*_k x_b\eta_b}{\sqrt{\psi_i^{*2} + \psi_k^{*2}}}\right). \quad (4)$$

Therefore, the conversions of the threshold and the two factor loadings between the old and new metrics are:

$$\frac{-\gamma_{\{i,k\}}}{\sqrt{\psi_i^2 + \psi_k^2}} = \frac{-\gamma^*_{\{i,k\}} + \lambda^*_i y_a - \lambda^*_k y_b}{\sqrt{\psi_i^{*2} + \psi_k^{*2}}} \quad (5)$$

$$\frac{\lambda_i}{\sqrt{\psi_i^2 + \psi_k^2}} = \frac{\lambda^*_i x_a}{\sqrt{\psi_i^{*2} + \psi_k^{*2}}} \quad (6)$$

$$\frac{\lambda_k}{\sqrt{\psi_i^2 + \psi_k^2}} = \frac{\lambda^*_k x_b}{\sqrt{\psi_i^{*2} + \psi_k^{*2}}} \quad (7)$$

Note that the unique variances provide essential scaling for thresholds and loadings pre- and post-transformation, but their own units are arbitrary. Because the models for the two instruments were fitted using identical unique variance identification constraints, the units for unique variances in the quad instrument model and the triplet instrument model are the same,

$$\psi_i^{*2} = \psi_i^2. \tag{8}$$

With this, Equations (5-7) simplify to:

$$-\gamma_{\{i,k\}} = -\gamma_{\{i,k\}}^{*} + \lambda_i^* y_a - \lambda_k^* y_b \tag{9}$$

$$\lambda_i = \lambda_i^* x_a \tag{10}$$

$$\lambda_k = \lambda_k^* x_b \tag{11}$$

With the transformation method determined, the next step was finding the equating coefficients x_a and y_a for each latent trait. The data structure called for a common-item non-equivalent group linking design (Kolen & Brennan, 2004, p. 19). Given the nested structure of the two instruments, all but 5 items with wording change could be used as common items, thus giving a high proportion of common items far exceeding the essential requirements. When equating, however, the common items are assumed to function in exactly the same way across instruments (Kolen & Brennan, 2004, p. 19). This assumption may not always hold in the present study, where contextual change across instruments takes place. However, the impact on the results due to possible violation of this assumption was expected to be small if the vast majority of items functioned in the same way across instruments. With this, the coefficients x_a and y_a were subsequently estimated by linear equating (Kolen & Brennan, 2004, p. 31):

$$\frac{\eta_a^* - mean(\eta_a^*)}{SD(\eta_a^*)} = \frac{\eta_a - mean(\eta_a)}{SD(\eta_a)} \tag{12}$$

where

- η_a denotes the latent trait in the default metric of the quad instrument model, thus $mean(\eta_a) = 0$ and $SD(\eta_a) = 1$ for the quad sample;

- η_a^* denotes the latent trait in a new metric, estimated by fitting a new model to the quad instrument sample, with all common item parameters fixed to values from the triplet instrument model, and $mean(\eta_a^*)$ and $SD(\eta_a^*)$ freely estimated.

With this, Equation (12) can be re-arranged:

$$\eta_a^* = SD(\eta_a^*)\eta_a + mean(\eta_a^*), \quad (13)$$

The linking coefficients x_a and y_a can thus be estimated:

$$x_a = SD(\eta_a^*) \quad (14)$$

$$y_a = mean(\eta_a^*) \quad (15)$$

The linking coefficients x_a and y_a for each of the 32 latent traits were thus obtained by extracting the latent $mean(\eta_a^*)$ and $SD(\eta_a^*)$ estimates from Mplus outputs. Given the large sample sizes and similar sample characteristics across instruments, the latent trait distributions were expected to be similar and it was therefore not surprising that most x_a coefficients were close to 1 and most y_a coefficients were close to zero (Table 3), with the deviations from the expected values reflecting differences between the two samples. The x_a parameters ranged from 0.782 to 1.016, indicating that the latent trait standard deviations of the quad sample were between 78% and 102% (i.e. generally smaller) of the triplet sample. One tentative explanation of such differences might be population change over time – perhaps the UK population from which operational assessment data were collected had become more diverse, thus explaining the variance increase from the older quad sample to the newer triplet sample. Another potential explanation might be demographic composition differences between the two samples. For example, there were a larger proportion of younger respondents in the triplet sample, which might explain why the "Rule Following" trait showed the largest variance increase. The item parameters for the quad instrument model

were then equated using these coefficients as shown in Equations (9-11) before subsequent analysis.

INSERT TABLE 3 NEAR HERE

Stability of item parameters. After equating, the item parameter sets were compared directly to establish their level of stability across the two instruments. The means and standard deviations of the differences and absolute differences were calculated. Note that the loading, threshold and unique variance parameters were scaled arbitrarily in accordance with the unique variance model identification constraints, and thus the size of the differences must be interpreted in line with the scaling of the parameters.

The relationships between parameter sets were also examined graphically using scatter plots. Multivariate outliers away from the equating line, which had standardized residuals of magnitude above 3, were identified and studied in the qualitative phase of the analysis.

Qualitative analysis of item context. Qualitative analysis of items was conducted for forced-choice blocks containing outliers as identified by the previous step of the analysis. To avoid confirmation bias, analysis was conducted purely though qualitative review of item text, without referring to their item parameter estimates. For each block concerned, analysis explored contextual changes across the triplet and quad versions of the block. Potential causes of parameter shifts were formulated, and predictions were made as to what the shifts may be. For a particular pairwise comparison of two items, contextual changes can cause parameter shifts in the following ways:

- When the context caused the likelihood of endorsement for one item over the other to change for the average person, the threshold is expected to shift;

- When the context moderated the relationships between items and their underlying traits, the loadings are expected to shift;
- When the context changed the amount of variation in the responses that cannot be explained by the underlying traits, the unique variances are expected to shift;
- When the context introduced sources of biases into the responding process, the existing model is insufficient for describing the full responding process, and all parameters can shift in unpredictable ways.

Themes emerging from qualitative analyses were reported in the Results section. Some general hypotheses of how the identified themes may influence item parameters in FC CAT are proposed in the Discussion.

Stability of trait score estimation. The ultimate goal of studying item parameter shift was to ensure stability of measurement at the trait level for each respondent. To assess this, respondents' scores based on parameter sets estimated from the two different instruments were compared. The sample taking the triplet instrument was selected for this analysis, because the binary outcomes of all pairwise comparisons were known in this sample. This sample was first scored using the parameters estimated from the triplet instrument, and then, separately, scored again using the before-equating parameters estimated from the quad instrument. Responses associated with the 5 items with wording change across instruments were not scored. At the end of this scoring process, each respondent in the sample had two sets of scores – one based on triplet instrument parameters, and the other based on quad instrument parameters. The trait scores estimated using the quad instrument parameters were then transformed using Equations (3) to align the metrics. The resulting two sets of trait score estimates were then compared as follows:

- Stability of rank ordering of individuals on a particular trait – correlations of the trait score estimates;

- Stability of rank ordering of individuals' personality profiles as a whole – correlations of profile locations (defined as the average score across all traits for each individual);
- Stability of rank ordering of traits for a particular individual – profile similarities (defined as the correlation between the two score profiles for the same individual based on two different parameter sets);
- Size of the differences between trait score estimates – relative and absolute differences between trait score estimates from different parameter sets.

Results

Stability of Item Parameters

Analysis was conducted on item parameter estimates that were neither associated with the 5 items with wording change, nor fixed in the model estimations. For example, there were 312 uniqueness terms in the models, one for each of the 312 items. However, 104 of them were fixed for model identification purposes and 5 were associated with items with wording change, thus reducing the total number of parameter estimates for analysis to 312 – 104 – 5 = 203.

The parameter estimates were aligned across the instruments, giving mean differences close to zero for all – thresholds, factor loadings and unique variances (Table 4). The parameters estimates also demonstrated strong linear relationships, as can be seen in the scatter plots of equated quad instrument parameters against triplet instrument parameters (Figures 1-3) and their very high correlations (Table 4). Estimates of item thresholds (see Figure 1) were mostly stable, giving a correlation of 0.975. Estimates of factor loadings (Figure 2) were less stable, giving a correlation of 0.878. Unique variance parameters turned out to be the most volatile to estimate across instruments (Figure 3), but still produced a high

correlation of 0.841. Regarding the spread of the estimates, while Figure 1 shows a uniform spread around the equating line for thresholds, Figure 2 shows clear heterogeneity in the spread of the factor loadings. Specifically, larger slopes varied much more between the instruments than smaller slopes did. The same was true for the uniquenesses (Figure 3). The greater fluctuations seen in loading and unique variance parameters were not surprising. Simulation studies by Brown and Maydeu-Olivares (2012; see Tables 3 and 4) showed that loading parameters were typically recovered less accurately than threshold parameters, with larger loading values providing greater space for fluctuations than smaller loading values. The uniqueness parameters were estimated with even less precision.

INSERT TABLE 4 AND FIRURES 1-3 NEAR HERE

Outliers. Between 2.0% and 2.6% outliers were identified for each type of parameter (Table 5). In total, 17 (5.5%) of the 307 common items (i.e., 312 items in the triplet version minus 5 items with wording change) were marked as outliers in at least one of the parameters. These outlier items were found in 8 (7.7%) of the 104 blocks.

INSERT TABLE 5 NEAR HERE

Qualitative Analysis of Item Context

Items within the 8 forced-choice blocks containing outliers were studied to identify contextual changes across the instruments. This analysis identified a number of recurring themes, which are outlined below and illustrated by examples.

Theme 1: Change in relative item endorsement levels. Change in relative item endorsement level was observed in 3 blocks. The block containing items 189-192 gives a good example.

Item		Quad	Triplet	Scale
189	I consider what motivates people	√	√	Behavioral
190	I am easily bored by repetitive work	√	√	Variety Seeking
191	I worry before an interview	√	√	Worrying
192	I finish things on time	√		Conscientious

The triplet version contains items 189, 190 and 191. In the workplace, item 189 is likely to be perceived as most desirable, so the relative endorsement levels of item 189 against items 190 and 191 are likely to be high. In the quad version, the desirability of item 192 is likely to be high. As a result, item 189 is no longer the obvious "best answer" in the quad, as it may be in the triplet. So the endorsement level of item 189 against items 190 and 191 is likely to be lower in the quad version. To put this in terms of parameters, the pairs {i189, i190} and {i189, i191} in the triplet version are likely to have lower threshold parameters (i.e., easier to endorse the first item) than in the quad version, which is what was observed.

Theme 2: Change in item's discrimination levels. Change in item discrimination levels was observed in 5 blocks. The block containing items 141-144 gives a good illustration.

Item		Quad	Triplet	Scale
141	I am lively in conversation	√		Outgoing
142	I follow rules and regulations	√	√	Rule Following
143	I persevere with tasks	√	√	Conscientious
144	I avoid talking about my successes	√	√	Modest

The triplet version contains items 142, 143 and 144, and it is clear to the respondent that they all refer to distinct attributes. The additional item 141 in the quad version, however,

is very similar to item 144 in content – both items have an element of talking to people. This "talking" emphasis in the same block creates an unintended contrast between items 141 and 144. As a result, item 144 may shift from being a positive indicator of Modest to being a negative indicator of Outgoing. Thus, the factor loading for item 144 on the Modest trait were expected to be lower in the quad version – exactly what was observed in the IRT parameter estimates. Predictions of shifts of other parameters in this block, however, were not as successful. It was hypothesized that item 142 would be unaffected by the shared "talking" element, and therefore the parameters for item 142 should not change. This prediction was not accurate and the loading for item 142 was actually lower in the quad version, suggesting that some additional factors were at play.

The qualitative study of change in context was unfortunately not always as simple as the examples given here. Often, multiple themes were present in the same block, leading to complex interactions and making the prediction of how item parameters would change extremely difficult. Nevertheless, based on this study, we suggested possible mechanisms behind some context-induced parameter shifts, which are summarized in the Discussion.

Stability of Trait Score Estimation

From a rank ordering perspective, trait score estimates for the same individuals based on different parameter sets were highly similar. Table 6 describes the correlations of scores for each of the 32 traits, the correlation of post-equating profile locations, and the profile similarities for all individuals in the sample (N=22,610). It was clear that the ordering of people at scale level as well as the similarity of whole personality profiles were preserved. The latter was important since selection decisions on comprehensive measures of personality were usually based on combinations of traits, not by comparing each individual trait.

INSERT TABLE 6 NEAR HERE

From an absolute difference perspective, the trait score estimates from different parameter sets were also highly similar. Table 6 describes the mean score differences and mean absolute score differences across the 32 latent traits. Reassuringly, most traits demonstrated mean differences close to zero and mean absolute differences of small magnitude. However, some traits demonstrated relatively large differences. The largest difference was seen in the "Conventional" trait, which showed mean difference of -0.183, suggesting that respondents typically received lower scores when scored using the quad instrument parameters as opposed to triplet instrument parameters. Note that one of the five items with wording change was from the Conventional trait and removed in the scoring process. The second largest difference was seen in the "Vigorous" trait, with mean difference of -0.164. Such differences may be caused by a combination of item parameter shift across instruments, item parameter estimation error and equating error.

Discussion

The parameter invariance assumption is fundamental to the full realization of adaptive personality assessments using the FC response format. The current study examined the effect of context on FC responding behavior, as represented by adding one extra item per FC block. Empirically-derived item parameters, estimated independently before and after the contextual change, were compared. The threshold, loading and unique variance parameters were largely stable. Furthermore, a small proportion (less than10%) of parameters that yielded substantial shifts, however, had little impact on the person parameter estimates. Evidence from the current study thus largely supported the parameter invariance assumption.

Nevertheless, a number of scenarios where this assumption was violated were reviewed, resulting in the identification of two recurring themes. The mechanisms behind parameter shifts are suggested below, and some recommendations for mitigating parameter shifts in adaptive FC assessments are made.

Themes in Influences of Context on FC Item Parameters

The two themes identified for parameter shifts are of particular interest to FC CAT implementations. Through understanding these themes better, appropriate test assembly rules can be designed to mitigate their occurrences, thus reducing the likelihood of parameter shifts and enhancing the accuracy of trait estimations. With this purpose in mind, we hypothesize possible mechanisms behind parameter shifts due to change in context for FC items.

Theme 1: change in relative item endorsement levels. In FC blocks, some items can appear more desirable than others, either because they are more socially appealing in general, or because they are more in line with the purpose of the assessment (e.g., Donovan et al., 2003; Kam, 2013; Paulhus & Vazire, 2007). When making comparative judgements in an assessment setting, respondents are likely to be considering the desirability of items consciously or unconsciously. As a result, when item desirability within a block is not balanced, endorsement can shift towards the more desirable "right answers".

There are several factors that may intensify such desirability-induced response biases. Firstly, it is likely to occur more often in high-stake situations, where respondents have stronger motivations to do well or appear good. Secondly, it is likely to be worsened when the desirability difference between items within the same FC block is large, thus making the perceived "right answer" more obvious to more respondents. Finally, it is likely to be more severe with smaller FC block sizes. In a block of two items, once the most desirable item is chosen to be "most" like the respondent, the other item has to be the "least" like the respondent, and the only information collected from this response is bias. But in a block of three items, the comparison between the remaining two items can still give useful information.

In terms of impact on measurement, such desirability-induced response biases introduce shifts in thresholds of the pairwise comparisons within the affected block, which can reduce the accuracy of latent trait estimation. To tackle this problem, items should be

worded neutrally or factually, so they do not sound obviously desirable or undesirable. Moreover, the relative endorsement levels of items should be estimated and controlled for in the instrument design. In a CAT setting, this translates into an additional rule in the test assembly algorithm – a numerical constraint preventing combinations of items with relative endorsement levels exceeding a certain acceptance threshold.

Theme 2: change in item discrimination levels. When considering several items simultaneously, respondents can perceive the item meaning differently to when they consider them independently. Most often, item interactions are caused by unplanned shared content between them, making their artificial similarity salient and deteriorating the original meaning of the items in relation to the attributes they indicate. Item interactions thus enhance or dilute the items' ability to measure their intended constructs, leading to shifts in item discrimination parameters.

There are several flags for identifying potential item interactions. The first clue comes from item wording – items sharing the same or synonymous keywords or phrases are likely to interact, as are items employing antonymous keywords. Furthermore, even if items do not explicitly share similar or opposite wordings, they can still have unplanned situational overlap that may lead to item interactions. The second clue comes from the constructs that the items measure – items from conceptually-similar constructs are more likely to interact than items from conceptually-distinct constructs.

In terms of measurement, item interactions can have two kinds of impacts. On one hand, when the shared context is not related to the latent constructs being measured, not only may the items have correlated residual variance caused by a common nuisance factor, but also do the items' focus shift towards that nuisance factor, thus reducing their power to measure the intended constructs. On the other hand, when the shared context is related to the latent constructs being measured, interaction-induced item cross-loading happens. In such

cases, the scoring model is no longer sufficient to model the response process. In a CAT setting, a viable solution to this problem is to prevent items that may interact from appearing in the same block. To do so, pairs of potentially interacting items need to be identified by subject matter experts and then coded in the test assembly algorithm as content "enemies" within (but not across) FC blocks.

Dealing with Change in Item Uniqueness

Unlike the case of item thresholds and loadings, parameter shifts in item unique variances are harder to explain and to predict. This is perhaps not at all surprising because unique variances are, by definition, residual variances unexplained by the responding model. Unique variances characterize how closely the actual item responses scatter around their predicted values. While unique variances reflect certain item properties, for example how central or peripheral the item is to the measured attribute, they may also depend on environmental factors external to the items that affect the level of random variation in respondents' answers.

In terms of measurement, less random variation in answers should reduce the residual variances of items and give more accurate trait estimates. While reducing residual variances is a good thing for measurement in general, there is one complication in a CAT setting – if the unique variance of an item changes, the parameter invariance assumption is violated. And because there is no simple way to precisely quantify the extent of random influences a priori, it is challenging to construct test assembly rules that standardize unique variances across blocks.

However, in practice, change in residual variances is less of a concern compared to shifts in other item parameters in FC CAT. In order for FC CAT to be effective, a large item bank with calibrated item parameters is required. While it is not too complicated to model a FC instrument with fixed block design as in the case of this study, it is impossible to calibrate

a large item bank using FC response data because of an astronomical number of combinations in which the items can be paired together. Therefore, large item banks designed for FC CAT assessments are calibrated using SS response formats, where the residual variances are likely to be at their highest due to many response biases that affect the SS format. Consequentially, the SS-based item parameters are likely to overestimate unique variances in FC CAT. This leads to overestimation of the resulting measurement error in FC CAT. The test assembly thus operates under a worst case scenario, making more conservative decisions regarding measurement precision, and arriving at more accurate trait estimates.

Unique variance fluctuations also have an impact on score estimation, through affecting the likelihood values of the responses. However, a small level of unique variance fluctuation is unlikely to dramatically change the score estimates. As can be seen in this study, an overall unique variance fluctuation characterized by a correlation of 0.841, together with a small number of shifts in other item parameters, still produced trait score estimates correlating to 0.991 or above. In summary, invariance of uniqueness in a FC CAT setting is given lower priority and importance compared to invariance of threshold or loading.

Limitations

One limitation of using historical data in this study is the confounding of contributions from contextual differences as well as potential sample differences in the observed parameter fluctuations. To partial out the contribution from potential sample differences, further studies need to incorporate adequate matching or randomization designs during data collection.

The contextual difference between the two instruments used in this study is also limited in nature. Firstly, both instruments were constructed manually by experts while taking into account content requirements and best practices in measurement, so the additional item seldom introduces significant contextual shift into a FC block. Once the human factor is

removed, computer-assembled FC blocks are likely to have larger impact of context, potentially leading to greater fluctuations in item parameters. Secondly, the FC block compositions were very similar across the two instruments, with three out of four items staying the same. The effect of fully shuffling the items into different blocks may lead to yet more contextual changes, and potentially larger item parameter shifts. This remains an area of research for further studies. However, the tight control over the context in this study is also its strength because it was possible to triangulate the potential causes behind the item parameter shifts, which would be much more difficult with less controlled contextual changes.

Finally, this study only focuses on measuring personality, which comprises relatively stable psychological constructs. For constructs that are more situation-dependent, contextual variations may lead to greater responding behavior differences. Therefore, generalizations of the findings in this study to FC assessments of other constructs must be made with caution.

Conclusion

An important assumption made by adaptive personality assessments using the FC response format is that the item parameters are invariant with respect to context. This study established the stability of item parameters across different FC personality assessments, and confirmed the robustness of person trait score estimation in the event of a small proportion of parameter shifts. Results from this study thus largely supported adopting the parameter invariance assumption. Furthermore, methods for preventing parameter shifts in FC CAT were suggested, with the aim of strengthening the parameter invariance assumption in practice for better measurement, and aiding future research on adaptive FC personality assessments.

References

Asparouhov, T., & Muthén, B. (2010). Multiple imputation with plus (version 2). Retrieved from https://www.statmodel.com/download/Imputations7.pdf

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1-26.

Bartram, D., Brown, A., Fleck, S., Inceoglu, I., & Ward, K. (2006). *OPQ32 technical manual*. Thames Ditton, UK: SHL.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment, 15*(3), 263-272. doi:10.1111/j.1468-2389.2007.00386.x

Brown, A., & Bartram, D. (2009). Doing less but getting more: Improving forced-choice measures with IRT. *Paper Presented at the 24th Annual Conference of the Society for Industrial and Organizational Psychology, 2-4 April,* New Orleans, LA. Retrieved from http://kar.kent.ac.uk/44788/

Brown, A., & Bartram, D. (2009-2011). *OPQ32r technical manual*. Surrey, UK: SHL Group.

Brown, A. (2012). Multidimensional CAT in non-cognitive assessments. *Paper Presented at the 8th Conference of the International Test Commission, 2-5 July,* Amsterdam, The Netherlands.

Brown, A. (2014). Item response models for forced-choice questionnaires: A common framework. *Psychometrika, Advance online publication.* doi:10.1007/s11336-014-9434-9

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational & Psychological Measurement, 71*(3), 460-502. doi:10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a thurstonian IRT model to forced-choice data using mplus. *Behavior Research Methods, 44*(4), 1135-1147. doi:10.3758/s13428-012-0217-x

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*(1), 36-52. doi:10.1037/a0030641

CEB. (2009-2014). *Global personality inventory - adaptive.* (Technical Manual). CEB.

CEB. (2014). *OPQ32r™ technical manual*. Surrey, UK: CEB.

Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance, 22*(2), 105-127. doi:10.1080/08959280902743303

Cheung, M. W. L., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(1), 55-77. doi:10.1207/S15328007SEM0901_4

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267-307. doi:10.1207/s15327043hup1803_4

Cornwell, J. M., & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to saville & willson (1991). *Journal of Occupational & Organizational Psychology, 67*(2), 89-100.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO–PI–R) and NEO five-factor inventory (NEO–FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement, 20*(4), 405-416. doi:10.1177/014662169602000407

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*(1), 81-106.

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the tailored adaptive personality assessment system (TAPAS) to support army selection and classification decisions (tech. rep. no. 1311)*. Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*(2), 155-166. doi:10.1037/0021-9010.84.2.155

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84-96.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? an examination of the frequency of applicant faking behavior. *Personnel Review, 36*(3), 341-355.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*(3), 167-184. doi:10.1037/h0029780

Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2008). Computerized adaptive testing of personality traits. *Zeitschrift Für Psychologie/Journal of Psychology, 216*(1), 12-21. doi:10.1027/0044-3409.216.1.12

Houston, J. S., Borman, W. C., Farmer, W. L., & Bearden, R. M. (2006). *Development of the navy computer adaptive personality scales (NCAPS).* (No. NPRST-TR-06-2). Millington, TN: Navy Personnel Research, Studies, and Technology Division, Bureau of Naval Personnel (NPRST/PERS-1).

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*(2), 153-162.

Kam, C. (2013). Probing item social desirability by correlating personality items with balanced inventory of desirable responding (BIDR): A validity examination. *Personality and Individual Differences, 54*(4), 513-518.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*(2), 312-320. doi:10.1037/0022-3514.55.2.312

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York, NY: Springer.

Lin, Y., Inceoglu, I., & Bartram, D. (2013). Towards creating forced-choice personality assessments ‘on the fly’: Do thurstonian IRT assumptions hold empirically? *Paper presented at the 78th Annual Meeting of the Psychometric Society,* Arnhem, The Netherlands.

McCrae, R. R., Costa Jr., P. T., & Martin, T. A. (2005). The NEO–PI–3: A more readable revised NEO personality inventory. *Journal of Personality Assessment, 84*(3), 261-270. doi:10.1207/s15327752jpa8403_05

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. I.,II. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88*(2), 348-355. doi:10.1037/0021-9010.88.2.348

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide (seventh edition)*. Los Angeles, CA: Muthén & Muthén.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15*(3), 263-280.

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology, 57*, 401-421. doi:10.1146/annurev.psych.57.102904.190127

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley & R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224-239). New York: Guilford.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, USA: Springer.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science (Wiley-Blackwell), 2*(4), 313-345. doi:10.1111/j.1745-6916.2007.00047.x

Robson, S. M., Jones, A., & Abraham, J. (2008). Personality, faking, and convergent validity: A warning concerning warning statements. *Human Performance, 21*, 89-106. doi:10.1080/08959280701522155

Stark, S., & Chernyshenko, O. S. (2007). Adaptive testing with the multi-unidimensional pairwise preference model. *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing,* Minneapolis, MN.

Stark, S., Chernyshenko, O., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*, 184-203.

Stark, S., & Chernyshenko, O. S. (2011). Computerized adaptive testing with the zinnes and griggs pairwise preference ideal point model. *International Journal of Testing, 11*(3), 231-247. doi:10.1080/15305058.2011.561459

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods, 15*(3), 463-487. doi:10.1177/1094428112444611

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81*(2), 332-342. doi:10.1037/0022-3514.81.2.332

Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology, 18*(5), 429-442.

van de Vijver, Fons J.R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage Publications.

van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35*(3), 346-360. doi:10.1177/0022022104264126

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*(2), 197-210. doi:10.1177/00131649921969802

Zavala, A. (1965). Development of the forced-choice rating scale technique. *Psychological Bulletin, 63*(2), 117-124. doi:10.1037/h0021567

Table 1

Sample composition

Sample characteristics		Quad instrument (OPQ32i)	Triplet instrument (OPQ32r)
Time of data collection		2004-2009	2009-2011
Gender	Male	62%	61%
	Female	38%	39%
	Missing	<1%	0%
Age	Below 20	1%	4%
	20-29	23%	33%
	30-39	32%	24%
	40-49	30%	21%
	50-59	12%	8%
	60 or above	1%	<1%
	Missing	1%	10%
Ethnicity	White	82%	56%
	Other	8%	8%
	Missing	10%	36%
N		**62,639**	**22,610**

Table 2

Stability of item parameter estimates across 10 imputations

Item parameter	Standard Deviation for item parameter estimates across imputations			
	Mean across all items	SD across all items	Min across all items	Max across all items
Threshold $\gamma_{\{i,k\}}$	0.007	0.009	0.001	0.079
Loading λ_i	0.008	0.007	0.001	0.051
Uniqueness ψ_i^2	0.013	0.024	0.000	0.206

Table 3

Equating coefficients

Latent trait (η_a)	x_a	y_a
1 Persuasive	0.929	-0.168
2 Controlling	0.908	-0.151
3 Outspoken	0.896	-0.111
4 Independent Minded	0.861	0.049
5 Outgoing	0.897	-0.043
6 Affiliative	0.852	-0.091
7 Socially Confident	0.831	-0.147
8 Modest	0.828	0.079
9 Democratic	1.016	-0.082
10 Caring	0.835	-0.233
11 Data Rational	0.819	-0.179
12 Evaluative	0.861	-0.257
13 Behavioural	0.910	-0.146
14 Conventional	0.901	-0.337
15 Conceptual	0.890	-0.196
16 Innovative	0.905	-0.258
17 Variety Seeking	0.830	0.033
18 Adaptable	0.841	0.031
19 Forward Thinking	0.884	-0.144
20 Detail Conscious	0.865	-0.298
21 Conscientious	0.864	-0.373
22 Rule Following	0.782	-0.358
23 Relaxed	0.921	-0.090
24 Worrying	0.809	0.085
25 Tough Minded	0.897	-0.147
26 Optimistic	0.885	-0.117
27 Trusting	0.807	-0.051
28 Emotionally Controlled	0.825	-0.057
29 Vigorous	0.785	-0.324
30 Competitive	0.952	-0.058
31 Achieving	0.886	-0.318
32 Decisive	0.896	0.030
Mean	**0.871**	**-0.138**

Table 4

Comparing item parameter sets estimated from quad and triplet instruments

	N	Quad		Triplet		Difference		Abs Difference		Correlation
Item parameter		Mean	SD	Mean	SD	Mean	SD	Mean	SD	
Threshold $\gamma_{\{i,k\}}$	302	-0.009	0.735	-0.028	0.751	0.019	0.167	0.121	0.116	0.975
Loading λ_i	307	0.731	0.290	0.726	0.330	0.005	0.158	0.104	0.118	0.878
Uniqueness ψ_i^2	203	0.484	0.459	0.497	0.737	-0.013	0.430	0.201	0.381	0.841

Table 5

Outliers with respect to parameter invariance from quad and triplet instruments

	Parameters			Affected Items			Affected Blocks		
	Total	Outlier	%	Total	Outlier	%	Total	Outlier	%
Threshold	302	7	2.3%	307	12	3.9%	104	5	4.8%
Loading	307	8	2.6%	307	8	2.6%	104	5	4.8%
Uniqueness	203	4	2.0%	307	4	1.3%	104	4	3.8%

Table 6

Comparing trait scores estimated using parameters from different instruments

	Mean	SD	Min	Max
Correlation of trait scores	0.996	0.002	0.991	0.999
Correlation of profile locations	0.985	-	-	-
Profile similarity	0.995	0.002	0.974	0.999
Mean score difference by trait	-0.088	0.041	-0.183	0.005
Mean absolute score difference by trait	0.113	0.031	0.050	0.184

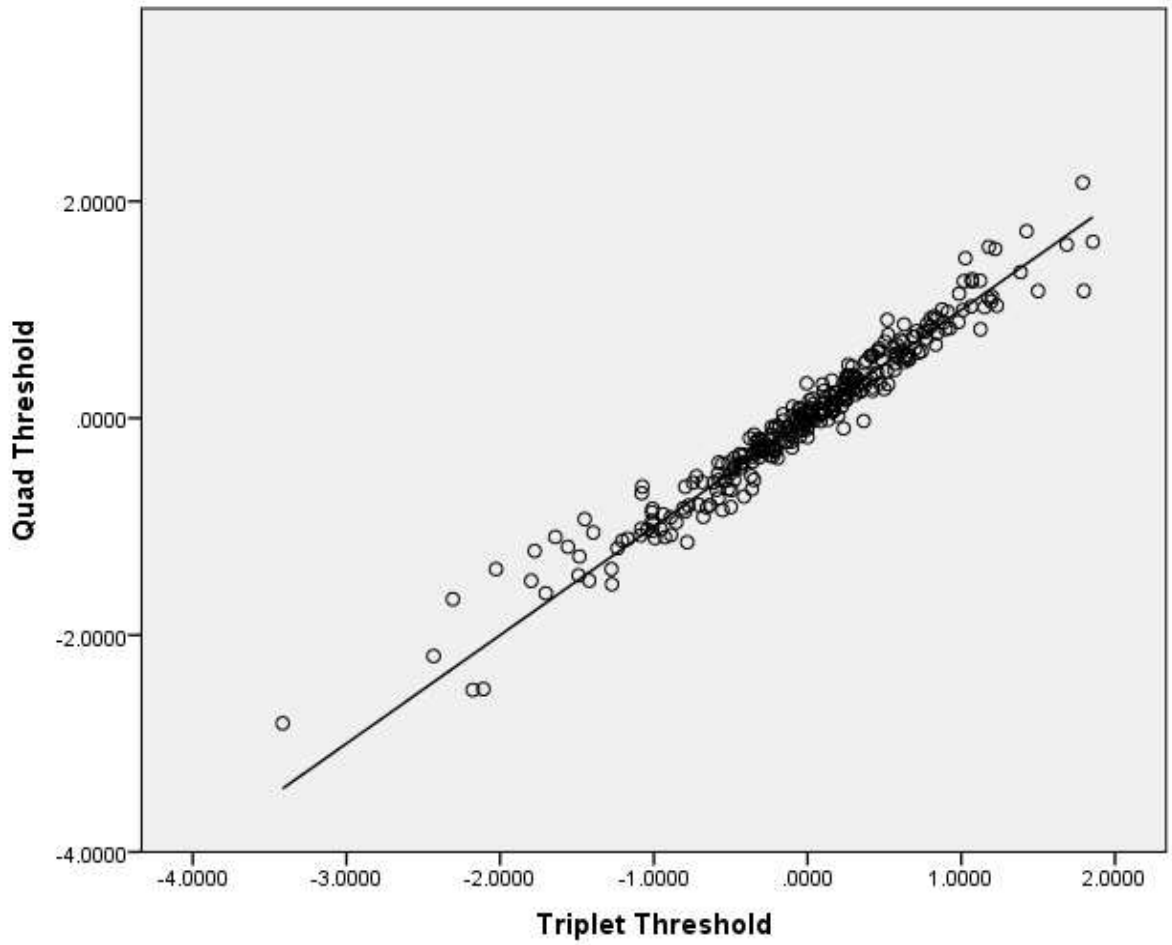

Figure 1. Scatter plot of estimated threshold parameters from quad and triplet instruments

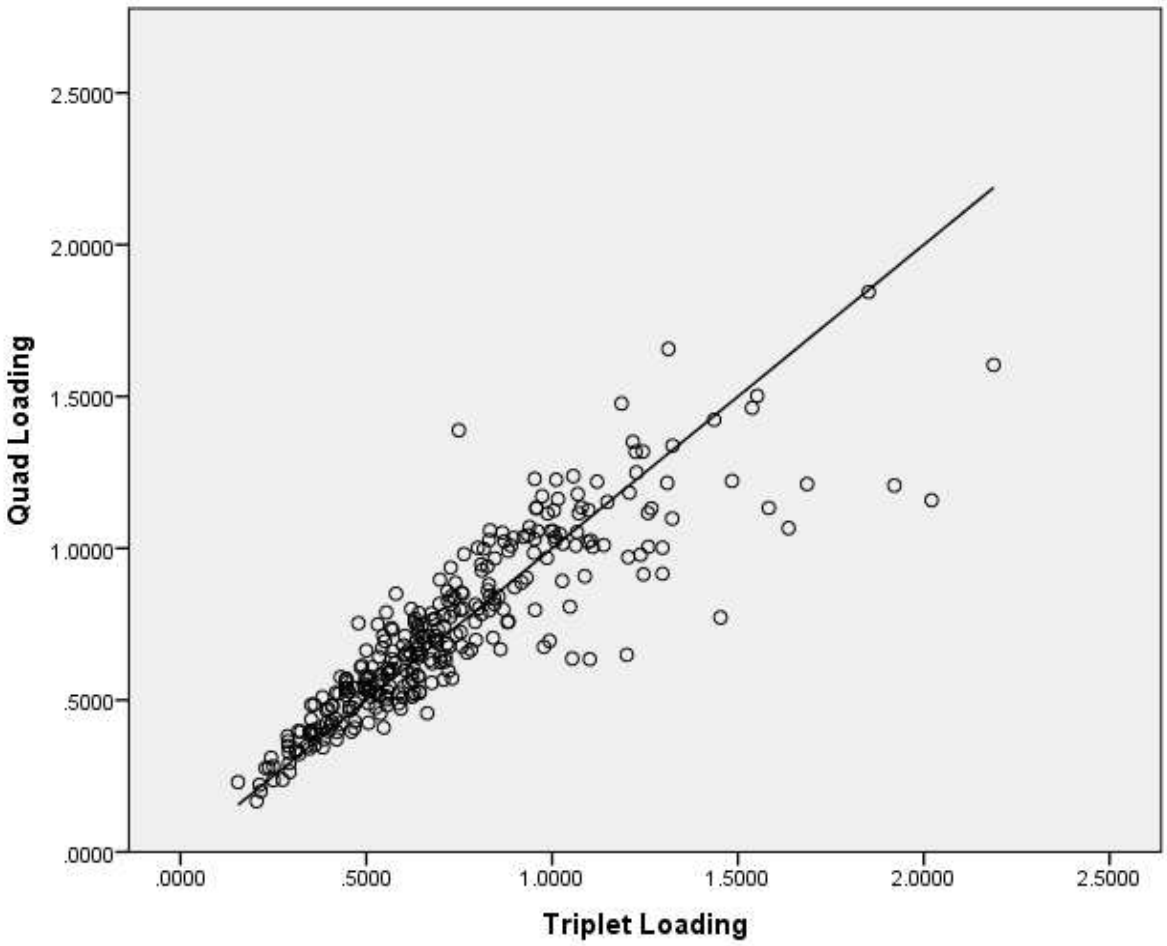

Figure 2. Scatter plot of estimated loading parameters from quad and triplet instruments

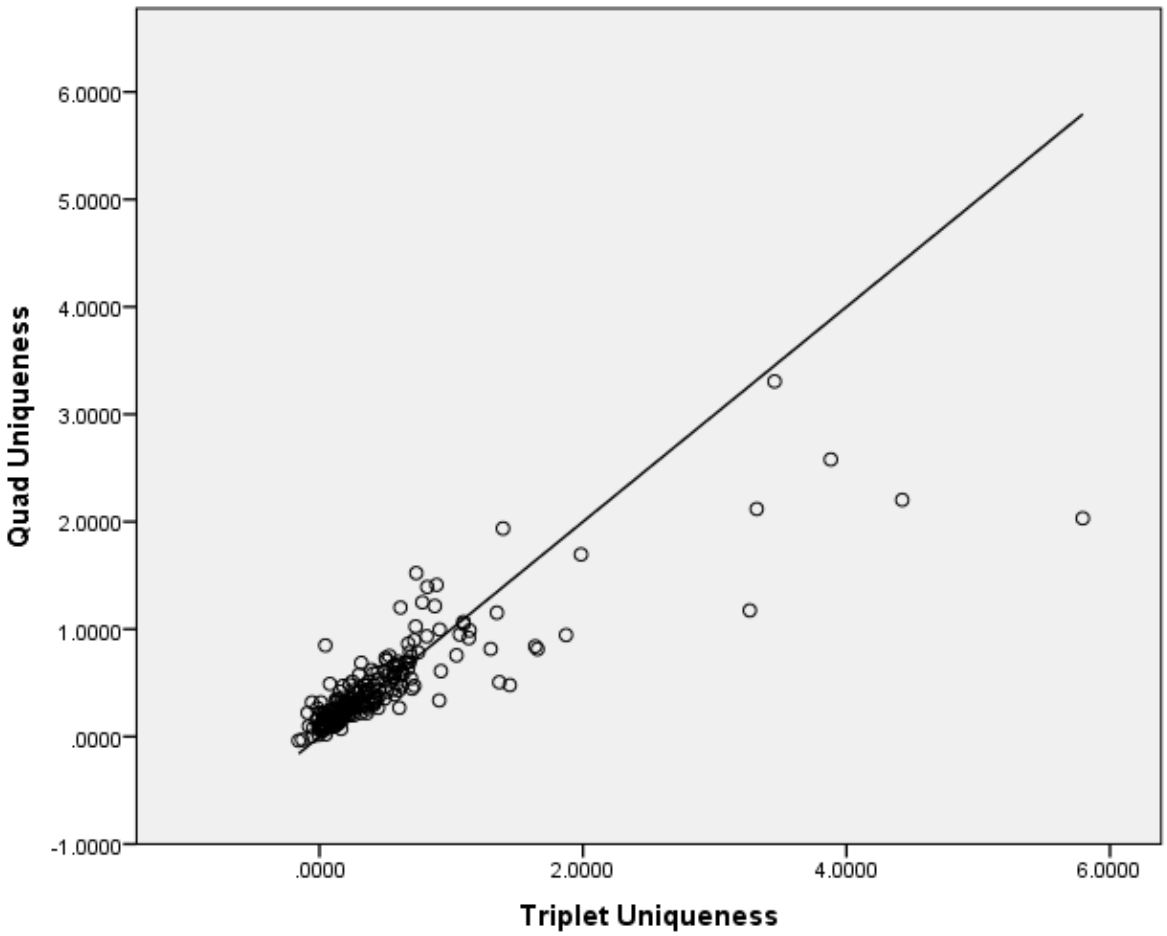

Figure 3. Scatter plot of estimated uniqueness parameters from quad and triplet instruments