Psychological Assessment
2012, Vol. 24, No. 3, 545–555


1040-3590/11/$12.00 DOI: 10.1037/a0026266

# Do Personality Scale Items Function Differently in People With High and Low IQ?

Chakadee Waiyavutti
Mahidol Wittayanusorn School

Wendy Johnson and Ian J. Deary
University of Edinburgh

Intelligence differences might contribute to true differences in personality traits. It is also possible that intelligence might contribute to differences in understanding and interpreting personality items. Previous studies have not distinguished clearly between these possibilities. Before it can be accepted that scale score differences actually reflect personality differences, personality items should show measurement invariance. The authors used item response theory to test measurement invariance in the five-factor model scales of the International Personality Item Pool (IPIP) and NEO-Five-Factor Inventory (NEO-FFI) across two groups of participants from the Lothian Birth Cohort 1936 with relatively low and high cognitive abilities. Each group consisted of 320 individuals, with equal numbers of men and women. The mean IQ difference of the groups was 21 points. It was found that the IPIP and NEO-FFI items were measurement invariant across all five scales, making it possible to conclude that any differences in IPIP and NEO-FFI scores between people with low and high cognitive abilities reflected personality trait differences.

*Keywords:* personality items, LBC1936, DIF, item response theory

Measurement invariance is an important property of psychological measures. Generally, we evaluate questionnaire or test responses by summing item responses, and we assess responses across groups of people who differ on some salient characteristics by comparing means and standard deviations. Then we interpret any differences by comparing these statistics to normative scores. In doing this, we implicitly assume that the questionnaire or test is measurement invariant (i.e., that it measures the same construct in the same way in different groups of people). When a questionnaire or test measures inconsistently across groups of people, mean differences in raw scores among the groups are confounded with other variables. In this case, measurement is not invariant, and comparisons of test scores cannot be made validly.

Several methods have been developed to test for measurement invariance. Item response theory (IRT; Lord, 1980) is one method that can be used in this way. IRT was originally developed to analyze scale properties of cognitive tests and has been widely used in educational assessments with high stakes, such as university admission. In these situations, it is very important that test items function equally across groups of examinees (i.e., the items are not biased for or against any particular group, so that test scores are comparable across the groups and the individuals within the groups). When cognitive abilities are measured, one should be confident that the score differences observed reflect true cognitive differences and are not measurement artifacts or explained by other variables. When this is not the case for some item on a test, we say that that item exhibits differential item functioning (DIF). For example, if people with low IQs interpret cognitive test items differently than those with high IQs, this will affect their probability of item endorsement for reasons unrelated to their actual cognitive abilities Although these issues have been in the forefront of high-stakes educational testing involving cognitive abilities, they are equally applicable to many other areas, including personality measurement.

IRT is currently one of the most powerful statistical tools to detect DIF, because a typical IRT model uses one single dimension to estimate both the trait levels of the respondents and the probabilities of endorsing the items, that is, the levels of the trait to which they refer. This use of a single dimension makes it possible to evaluate simultaneously how groups perform on a test and whether items function similarly for people in different groups. For example, Drasgow (1987) used IRT to examine DIF across sexes and races in the English and Mathematics sections of the American College Test and found measurement invariance across Hispanic, Black, and White men and women. In contrast, more recently, Teresi et al. (1995) found considerable DIF on six cognitive screening measures across African American, Latino, and White non-Latino groups, as well as groups with low and high levels of education. For example, they found that some items were less relevant to cognitive ability in African American and Hispanic

This article was published Online First November 14, 2011.

Chakadee Waiyavutti, Mahidol Wittayanusorn School, Nakorn-Pathom, Thailand; Wendy Johnson and Ian J. Deary, MRC Centre for Cognitive Ageing and Cognitive Epidemiology and Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom.

This work was supported by the University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology. The Centre is funded by the Biotechnology and Biological Sciences Research Council, the Engineering and Physical Sciences Research Council, the Economic and Social Research Council, and Medical Research Council Grant G0700704 as part of the Lifelong Health and Wellbeing Initiative. Wendy Johnson was supported by Research Council of the United Kingdom Fellowship #GR/T27983/01. The University of Edinburgh is a charitable body, registered in Scotland with registration number SC005336.

Correspondence concerning this article should be addressed to Chakadee Waiyavutti, Mahidol Wittayanusorn School, Nakorn-Pathom, Thailand 73170. E-mail: chakadeew@hotmail.com

groups than in White groups and for those with low, compared with high, education. Moreover, some items were answered correctly more frequently by White people than by African Americans with otherwise similar levels of cognitive ability. Similarly, Crane et al. (2006) found DIF in at least four out of 10 items on the Italian Mini-Mental State Examination across age and educational attainment groups.

IRT has also been used to investigate DIF in personality scales. Because personality scales often use categorically ordered multiple responses, rather than the binary (right/wrong) responses often used in cognitive tests, more complex models are required. Generally, personality items require respondents to rate the degree to which items describe them along an ordered, graduated scale; for example from 5 (*very characteristic of me*) to 0 (*very uncharacteristic of me*). The concepts involved, however, are the same. For example, Smith and Reise (1998) observed that women scored about 0.4 standard deviations higher than did men on the stress reaction scale of the Multidimensional Personality Questionnaire (MPQ; Tellegen & Waller, 2008). Using IRT, Reise and Smith found numerous items that displayed DIF. Items involving vulnerability were more likely to be endorsed by women, and some items involving tension were more likely to be endorsed by men, despite apparently comparable levels of stress reaction. However, sex differences in scale scores without the DIF items remained unchanged. They concluded that the stress reaction items indeed reflected valid mean differences in stress reaction scores between men and women, despite the apparently differently functioning items.

Soon after that, Reise, Smith, and Furr (2001) again used IRT to investigate gender differences, this time on the neuroticism scale of the NEO-PI-R (Costa & McCrae, 1992), a popular measure highly correlated with other measures of negative emotionality, such as the Stress Reaction scale of the MPQ. The anxiety facet of the neuroticism scale also had consistently shown higher female than male scores by about 0.25 standard deviations. Overall, results indicated that anxiety items of the NEO-PI-R measured the same construct for men and women, as item loadings on the latent neuroticism factor could be constrained to be equal. However, many anxiety items displayed DIF, indicating that they measured differently in men and women; some items increased male scores, and some increased female scores. Similar to the findings in their previous study (Smith & Reise, 1998), these differences were influenced by between-item relations and specific gender interactions with item content. For example, fear items were endorsed more frequently by women, and worry items were endorsed more frequently by men, but mean gender differences on fear items were more likely to be larger than those on worry items. Therefore, women tended to have higher scores on the anxiety scale without actually reflecting higher anxiety.

IRT has also been used to detect DIF in personality items across cultures, as cultural differences might influence mean differences in personality scales. For example, observing mean differences on MPQ scales across samples matched on age and gender from Minnesota and Germany, Johnson, Spinath, Krueger, Angleitner, and Riemann (2008) found that many items of the MPQ showed differential function in the two samples. These differences accounted for mean differences between the two cultures on most MPQ scales, but without preexisting theory, it was not known if these DIF items truly reflected either cultural differences or translation difficulties. Detecting DIF items in this study, however, was valuable in the sense that it indicated that these items were equally relevant to the scale constructs in the two samples, but they were differently relevant to the levels of the scale constructs in the two cultures. For example, Germans tended to endorse well-being items involving mood more than did Minnesotans, who tended to endorse items involving quality of experience more than did Germans. These studies indicate the practical usefulness of IRT in DIF detection in personality scales across people from different backgrounds.

Cognitive ability is a strong contributor to individual behaviors. Some have suggested that it is necessary to study personality constructs in conjunction with cognitive constructs (e.g., Sternberg & Ruzgis, 1994) if either one is to be understood properly, because completion of personality scales requires reading skill and the ability to evaluate one's own personality and behavior in relation to those of others. Although personality scale authors attempt to keep reading at the level at which people in most developed countries are allowed to leave school (i.e., around 8th grade), it is possible that people with different levels of cognitive ability tend to think about and manifest their otherwise similar personalities in different ways that impact how they complete personality scales.

Several studies have investigated associations between personality and cognitive ability. To date, these studies have not been successful in clarifying the associations, because potential differences in measurement properties of scales in people with different levels of intelligence and personality have been confounded with the potential personality-intelligence associations. For example, Austin, Deary, and Gibson (1997) tried to assess whether the NEO-FFI scales were measuring in the same way in two groups of Scottish farmers with different levels of IQ by comparing descriptive statistics between the groups. Although they found that people with high IQs had higher Openness to Experience scores and lower Neuroticism scores than those with lower IQs, they also found low reliability in Openness to Experience scores and more differentiation in Neuroticism scores for people with low IQs. Because group differences were confounded with measurement differences in these statistics, they could not conclude that the differences actually represented differences in Openness to Experience or Neuroticism.

More recently, Austin et al., (2002) again investigated correlations between personality and cognitive ability scores. They found that people with low IQs tended to have higher Psychoticism and Neuroticism scores, as measured by Eysenck's scales. They realized, however, that those relations could have occurred because people with low IQs could not differentiate items intended to reflect Psychoticism from items intended to reflect Neuroticism. Thus, they suggested that their results were valid only if the measures they used were equivalent at the item level. Toomela (2003) used facet-level analysis to study the factor structure of the scales of the NEO-PI-R in people with low and high cognitive abilities. His results suggested that the NEO's facet items could not equally differentiate people with differences in cognitive abilities. Because he was working at the facet level, however, Toomela (2003) could not specify which items in each facet provided different discriminating power across the groups. In a related vein, Rammstedt, Goldberg, and Borg (2010) used factor analysis to study the Big-Five factor structure of personality in people with different levels of education. They found that the five-factor struc-

ture held well only in subsamples of highly educated persons. However, after controlling for acquiescence bias, the Big-Five structure emerged overall.

Taken together, research to date leaves it unclear whether personality items can measure people with low and high cognitive abilities equivalently, as well as whether there are differences in personality associated with differences in cognitive ability. IRT provides a powerful tool with which to distinguish measurement equivalence from group differences, and we used it in this study to make this distinction. To do this, we made use of the International Personality Item Pool (IPIP; Goldberg, 1999) and the NEO-Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1992), two inventories based on the Five-Factor model of personality traits, in groups of individuals with low and high IQs from a sample of healthy older people very homogeneous for age and ethnic background. After testing measurement invariance, we investigated mean differences between the two IQ levels in the personality scale scores on the invariant items and compared the results to those of previous studies that failed to test for measurement invariance.

## Method

### Participants

The data were obtained from the Lothian Birth Cohort 1936 (LBC1936; Deary et al., 2007). The intent of the LBC1936 study was to follow up on the life outcomes of some of the participants in The Scottish Mental Survey 1947 (SMS1947) in older age. The SMS1947 was conducted on June 4, 1947. On that date, 70,805 children who had been born in 1936 (94.14% of the population) and were attending school anywhere in Scotland were administered a version of the Moray House Test No. 12 (MHT), a group intelligence test. As follow-up of all SMS1947 participants was not practical, participants residing in the region of the city of Edinburgh (known as the Lothian area) were targeted. LBC1936 recruitment began in 2003 with permission from the Lothian Health Board to identify potential participants of that original survey (SMS1947) who were living in the Lothian area immediately surrounding the city of Edinburgh at that time, to obtain a subsample of the original survey population representative of those now living in one geographic area within Scotland. This was done using the Community Health Index, which is a list of residents who live in an area and are registered with a general medical practitioner. On this list, there were 3,810 individuals who had been born in 1936. During 2004–2006, 3,686 of them were located and invited by mail to take part in the study. Of this group, 1,132 were interested and eligible to take part. Eligible individuals were participants in SMS1947, in good health, and able to attend the assessment protocol. A second mailing generated an additional 94 who were interested and eligible to participate. Of these 1,226 individuals, 85 withdrew from the study and 50 could not be contacted to make test appointments before the end of testing in May 2007.

In total, 1,091 participated in LBC1936. They were about the age of 70 years when Deary et al. (2007) conducted the assessment, which consisted of a physical examination and an interview, a broad range of cognitive ability tests, including the MHT, and completion of study questionnaires about lifestyle, personality, demographic characteristics, life history, and other variables. The personality inventories were the well-validated 50-item version of the IPIP (Goldberg, 1999) and 60-item version of the NEO-FFI (Costa & McCrae, 1992). For this study, we used the age-70 MHT scores to measure IQ and the scale scores of the IPIP (Goldberg, 1999) and NEO-FFI (Costa & McCrae, 1992) questionnaires to measure personality.

Of the 1,091 participants, 12 were excluded because they did not complete the MHT at age 70. In addition, 55 participants had MHT scores that translated to IQ-scale scores lower than 70 within the LBC sample. We eliminated these participants because it seemed likely that their cognitive function might have made it difficult for them to complete the IPIP and NEO-FFI. An additional 144 participants were removed because they did not complete the IPIP and NEO-FFI at age 70. Therefore, we had 880 participants who had completed the IPIP and NEO-FFI and had MHT scores ranging from 70 to 123. We used the extreme-group approach to assess whether the IPIP and NEO-FFI items functioned differently in people with relatively low and high cognitive abilities. The group with low IQ had IQ-scale MHT scores ranging from 70 to 100 within the LBC sample, with a mean of 91 ($SD$ = 7.6), and the group with high IQ had scores ranging from 108 to 123, with a mean of 113 ($SD$ = 3.5). Thus, an additional 193 participants with IQ-scale scores ranging from 101 to 107 were removed. The mean IQs in the groups differed by 21 points ($SD$ = 12.6), with a large effect size (Cohen's $d$ [mean difference/pooled standard deviation] = 1.67). Cohen (1992) suggested that effect sizes of 0.20 are small, 0.50 are moderate, and 0.80 are large. There were no significant differences in mean IQs between men and women overall, nor were there any significant mean differences between men and women in the low-IQ group, nor between men and women in the high-IQ group.

The groups were kept equal in total size and numbers of men and women to minimize the possibility that any differential item function we detected was the result of demographic differences in the groups. Ten participants in the selected range—three men in the group with high IQs and four men and three women in the group with low IQs—were randomly selected for exclusion to maintain the same numbers in each group. Theoretically, item parameters from IRT models are robust to variations in sample characteristics, but practically speaking, this is not always the case (Embretson & Reise, 2000).

### Measures

**Moray House Test No.12 (MHT).** The MHT is a well-validated general cognitive ability test, often considered a verbal reasoning test (Deary, Whalley, & Starr, 2009). It was originally designed to recruit students to grammar school education. The highest possible score is 76 from 71 items.

The test was administered individually to the participants with a 45-min time limit. As noted previously, on June 4, 1947, all children who were born in 1936 and attending schools in Scotland were administered the MHT to assess their cognitive abilities. The mean raw score on the test in the full population was 36.6 ($SD$ = 15.8, range = 1–74). The mean raw score of children at the age of 11 years in 1947, who later became participants of LBC1936, was 49.0 ($SD$ = 11.8). The mean raw score of LBC 1936 participants at about age 70 was 64.2 ($SD$ = 8.8, range =

9–76). Our participants had higher cognitive abilities than the overall population, at least partly because they represented the proportion of the population that had survived to age 70 (Deary et al., 2007). This is of little relevance to our study, however, as we were able to select two groups of participants that showed marked differences between those classed as of high and low IQ at age 70. The scores for the full LBC1936 sample were adjusted for age at testing and converted to the standard IQ scale with mean of 100 ($SD = 15$).

**International Personality Item Pool (IPIP).** The IPIP is a broad-bandwidth personality inventory. It is not in commercial use as it was developed to serve personality researchers. The version used in this study contains 50 items. There are 10 items measuring each of five personality scales. The scales are known as the Big-Five personality factors and include Extraversion (E), Agreeableness (A), Conscientiousness (C), Emotional Stability (ES), and Intellect (I). Generally, IPIP items consist of phrases describing common behaviors or emotional experiences. An example is "feel little concern for others." Participants were asked to read through the items and rate from 1 (*very inaccurate*) to 5 (*very accurate*) how accurately the phrases described them when compared with other people of the same sex and similar age.

**NEO-Five-Factor Inventory (NEO-FFI).** This is a 60-item inventory. It is a short version of NEO-PI-R, measuring five scales which are analogous to the IPIP, except the ES scale in the IPIP is reversed as Neuroticism (N) in the NEO-FFI. The I scale in the IPIP is comparable to Openness (O) in the NEO-FFI. Participants rated each item on the 5-point Likert-type rating scale from 0 (*strongly disagree*) to 5 (*strongly agree*).

For the NEO-FFI, there were 11 items for which some response options were not endorsed at all. Unendorsed response options included "strongly agree," "agree," "disagree," and "strongly disagree" but not the neutral options. Women with high IQs did not endorse the "strongly disagree" options of Items 4 and 5. The IRT models could not be run when this was the situation, but we wished to retain all items. It was thus necessary to collapse the NEO items' five response options to three by combining Response Options 1 and 2, keeping Response Option 3 (a neutral response) as the original, and combining Response Options 4 and 5. We might have considered dichotomizing these items, but this could not be done clearly because the neutral option was always endorsed. Moreover, we wished to retain as much of the original response format as we could. As a result, our IRT models for all NEO-FFI items had three NEO-FFI response options ranging from Response Option 1, indicating disagreement, to Response Option 3, indicating agreement.

On all scale scores of the IPIP and NEO-FFI, the low-IQ group had lower reliabilities than did the high-IQ groups. On the IPIP, the reliabilities of Intellect scores were the lowest (.66 for the low-IQ group and .79 for the high-IQ group), and those of Emotional Stability scores were the highest (.86 for the low-IQ group and .89 for the high-IQ group). On the NEO-FFI, the high-IQ group had the lowest reliability for Openness scores (.75) and the highest reliability for Neuroticism scores (.88), whereas the low-IQ group had the lowest reliability for Agreeableness scores (.59) and the highest reliability for Conscientiousness scores (.84).

## IRT Analysis

Item response theory (IRT) defines a mathematical model of an individual's probability of responding to an item as an indication of the individual's level of the trait dimension tapped by the item. Under the model, parameters representing both the items comprising a scale and the individuals comprising a sample are measured along a single latent trait dimension. People parameters refer to the locations of the individuals along the trait dimension. They were not the focus of interest in this study, however, so we discuss them no further. The most commonly used IRT model generates two parameters for each item, known as difficulty and discrimination.

When IRT was first developed, it was applied to intelligence, ability, and achievement test items, which have responses that can be considered correct or incorrect. Some of the vocabulary used to describe the item properties of interest reflect this initial application and are not intuitively descriptive when applied to personality items that do not have correct and incorrect responses. The original usage persists in the literature, however, so we follow it here. Item difficulty refers to the point along the trait dimension where a person with that level of the trait has a 50% probability of endorsing the item. Personality items with higher difficulties tend to be endorsed only by those who are higher on the trait in question. Item discrimination refers to the degree to which endorsement of an item actually reflects the endorsing individual's level on the trait dimension associated with the item (i.e., how good the item is at measuring what it is supposed to measure at a particular level of difficulty). The standard IRT model estimating item difficulty and discrimination can be expressed by the following equation:

$$P_i(\theta_j) = \{1 + exp[1.7\alpha_i(\theta_j - \beta_i)]\}^{-1}$$

This equation indicates that respondent $j$'s probability of endorsing item $I$, $P_i(\theta_j)$, is a logistic function of the respondent's trait level ($\theta_j$), the item difficulty parameter ($\beta_i$), and the item discrimination parameter ($\alpha_i$).

Generally, personality measures use ordered category responses, with several options, rather than a simple dichotomous yes/no response. This complicates the IRT model. Essentially, it means that each item must have difficulty and discrimination parameters for the boundaries between all response option categories. Under this more complex model, the interpretation of these parameters remains the same, however. For example, the difficulty parameter for the lowest response option represents the trait level at which a respondent has a 50% probability of moving from that lowest response option to the next lowest response option, and the discrimination parameter represents how good that response option boundary is at capturing a real difference in the respondent's level of the trait. IRT models rely on the assumptions of unidimensionality and local independence. Unidimensionality means that all items can be considered to measure a single trait and only that single trait. Local independence means that, controlling for a respondent's trait level and item parameters, the probability of endorsing one item is independent of the probability of endorsing any other. With these assumptions, the IRT model accounts completely for the coherence of scales. We tested whether IPIP and NEO-FFI items met the assumption of unidimensionality by separately factor analyzing the items for each scale in each IQ group to see whether there were single dominant factors running through the items. The ratios for the first to second eigenvalues for the E,

A, C, ES, and I scales of the IPIP were, respectively, 2.62, 2.66, 2.06, 4.04, and 1.92 for the low-IQ group and 4.83, 4.46, 3.40, 5.09, and 2.80 for the high-IQ group. For the N, E, O, A, and C scales of the NEO-FFI. The ratios were, respectively, 3.60, 2.40, 2.09, 1.92, and 3.53 for the low-IQ group and 4.13, 2.61, 2.72, 3.12, and 4.04 for the high-IQ group. These ratios were typical of those found in personality scales to which IRT has been applied in other studies and suggested that the scales had sufficient unidimensionality in both groups to proceed with the IRT analysis.

## Investigating Differential Item Functioning

A common reason for examining scale performance of different groups of people is that it is possible that the groups may show differences in means and/or variances in the trait in question. However we cannot be sure that these differences are real unless we are clear that the scale measures the same construct in the same way in different groups. IRT provides a way to check whether this is the case. Under IRT models, items may function differently in different groups of people if either the item difficulty or the item discrimination parameters, or both, differ between the groups. If the items do function differently, we cannot interpret mean and variance differences in scale scores from the items in any direct way. Only if the items function the same way in different groups can we take mean and variance differences in scale scores between the groups at face value. In our study, we tested whether items functioned in the same way in the IRT models by first estimating the item parameters for each scale of the IPIP and the NEO-FFI separately in groups with relatively low and high IQs and evaluating the IRT model fit. We then constrained all the item parameters equal across the two groups and observed whether this constrained version of the model fit substantially less well. If it did, then it was necessary to conclude that there must be some differences in item function. In that event, additional work was necessary to determine exactly what the differences were and why they existed. If it did not (i.e., if the constrained model fit equally well) then we could conclude that the items functioned the same way in the two groups.

To evaluate model fit, we computed differences in $-2^{*}$log-likelihood between models. These are distributed as chi-squared, with degrees of freedom equal to the estimated parameters. Small and nonsignificant chi-square differences indicate that models fit equivalently, and more constrained models are then preferred for reasons of parsimony. Because chi-square differences tend to show significance purely due to sample size, we also considered the information theoretic fit statistics Bayesian information criterion (BIC; Schwarz, 1978) and Akaike information criterion (AIC; Akaike, 1983) to evaluate our models. These model fit statistics give preference to more parsimonious models and are less sensitive to sample size. Smaller values indicate preferred models. Practically, we focused primarily on BIC, because it has tended to be more accurate in recovering the true model in simulation studies (Markon & Krueger, 2004).

The IRT models were implemented in the software Mplus (Muthén & Muthén, 1998–2007), providing multiple group analysis with ordered categorical data using maximum likelihood estimated with robust standard errors.

Where we could establish measurement invariance, we also explored associations between personality and cognitive ability. To do so, we used analysis of variance to test for statistical significance of mean differences in IPIP and NEO-FFI scores between people with relatively low and high IQs, as well as between people with different sex but similar IQs. To allow for the many statistical tests we anticipated running, we reported significance levels of $p < .01$ and $p < .001$.

## Results

Differences in mean responses of people with low and high IQs were significant across all scales of the IPIP and NEO-FFI ($p < .01$). People with high IQs scored significantly higher than those with low IQs with a moderate effect size (Cohen's $d = 0.21$) for the Intellect scale, and with small effect sizes ($d = 0.15$ and .14) for the Emotional Stability scale of the IPIP and Openness to Experience scale of the NEO-FFI. The effect sizes of differences ranged from .02 to .07 for all the other scales of both measures except the Neuroticism scale of the NEO-FFI. On this scale, people with low IQs scored higher than those with high IQs with a small effect size ($d = 0.18$). These significant differences in mean response rates for each scale of both measures represented differences in the raw scale scores of people with different cognitive abilities and sexes. On both the IPIP and the NEO-FFI, people with relatively low IQs used options indicating agreement with items (Options 1 and 2) more but used options indicating lack of agreement with items (Options 4 and 5) less than those with higher IQs across all scales, but they used Option 3 at about the same rates. Moreover, women scored significantly higher than men on the Agreeableness scale of both measures and on the Neuroticism scale of the NEO-FFI, with moderate effect sizes (Cohen's $d = 0.46$, .30, and .21, respectively), and on the Openness to Experience scale of the NEO-FFI, with a small effect size ($d = 0.18$). Men scored significantly higher than women only on the Emotional Stability scale of the IPIP with a small effect size ($d = 0.12$). Men and women showed similar patterns of responses to items from the Extraversion and Conscientiousness scales. However, it appeared that men used Options 4 and 5 on items from the Agreeableness and Emotional Stability scales of the IPIP and the Neuroticism scale of the NEO-FFI scales more often than did women. Distribution frequencies showed that respondents tended to endorse options indicating agreement with the items over options indicating lack of agreement ($p < .001$). Further details of observed responses and response rates to the IPIP and NEO-FFI items in all scales across groups of people separately are available from the first author on request.

Each model we fit consisted of all the items from particular IPIP and NEO-FFI scales in either men and women of any IQs, people of either sex with low and high IQs, or men with low and high IQs or women with low and high IQs. Thus, we fit models including all the items from the Agreeableness scale in each of these combinations, all the items from the Conscientiousness scale in each of these combinations, and so on for the rest of the IPIP and NEO-FFI scales. The constrained models for an example item from the IPIP are shown in Table 1 and an example item from the NEO-FFI in Table 2. Table 1 shows Item 1 from the Extraversion scale of IPIP Item 1, "I am the life of the party," and Table 2 showed the other from NEO-FFI Item 2, "I like to have a lot of people around me," in groups with low and high IQs of each sex, under the various forms of model constraints applied. For each model of each item,

Table 1
*Parameter Estimates of an Example Item (IPIP Item 1, "I Am the Life of the Party") From the Extraversion Scale of the IPIP Across IQ and Gender Groups*

| | Men | | | | | | | | | | Women | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Low IQ | | | | | High IQ | | | | | Low IQ | | | | | High IQ | | | | |
| Parameters constrained | α | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | α | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | α | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | α | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ |
| None | 1.16 | −2.66 | −1.91 | 0.14 | 2.29 | 1.98 | −2.32 | −0.67 | 1.95 | 6.26 | 1.58 | −2.93 | −2.13 | 0.65 | 3.34 | 1.52 | −2.24 | −0.73 | 1.96 | 4.89 |
| All | 1.43 | −1.98 | −0.86 | 1.36 | 4.07 | | | | | | 1.54 | −2.06 | −0.92 | 1.80 | 4.60 | | | | | |
| β | 1.28 | −2.10 | −0.98 | 1.24 | 3.93 | 1.55 | | | | | 1.64 | −2.05 | −0.92 | 1.80 | 4.60 | 1.43 | | | | |
| A | 1.57 | −2.91 | −2.12 | 0.03 | 2.32 | | −2.17 | −0.61 | 1.86 | 5.83 | 1.55 | −2.93 | −2.93 | 0.59 | 3.25 | | −2.28 | −0.76 | 2.00 | 4.98 |

| | Low IQ | | | | | | | | | | High IQ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Men | | | | | Women | | | | | Men | | | | | Women | | | | |
| Parameters constrained | α | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | α | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | α | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | α | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ |
| None | 1.16 | −2.67 | −1.91 | 0.13 | 2.28 | 1.58 | −1.80 | −0.99 | 1.78 | 4.47 | 1.98 | −3.46 | −1.81 | 0.81 | 5.12 | 1.52 | −2.25 | −0.74 | 1.96 | 4.89 |
| All | 1.29 | −1.79 | −1.03 | 1.30 | 3.61 | | | | | | 1.75 | −2.39 | −0.81 | 1.86 | 5.39 | | | | | |
| β | 1.24 | −1.73 | −0.96 | 1.37 | 3.67 | 1.33 | | | | | 1.89 | −2.50 | −0.92 | 1.74 | 5.29 | 1.59 | | | | |
| A | 1.35 | −2.88 | −2.10 | 0.01 | 2.25 | | −1.73 | −0.95 | 1.70 | 4.27 | 1.76 | −3.41 | −1.82 | 0.72 | 4.84 | | −2.32 | −0.77 | 2.04 | 5.07 |

*Note.* IPIP = International Personality Item Pool; β = difficulty parameter; α = discrimination parameter; Low = Low-IQ group (range = 70–100); High = High-IQ group (range = 108–123).

we estimated one discrimination parameter (α). High values of α indicated that items measuring a particular trait were strongly related to that trait and that they worked well in capturing real differences in respondents' levels of the trait. Each model displayed four ($\beta_1 - \beta_4$) and two ($\beta_1 - \beta_2$) difficulty parameters for the IPIP and NEO-FFI items, respectively (because we preserved all five response options of the IPIP's items and reduced the NEO-FFI's items' responses to three), in an ascending order indicating that lower categorical options reflected lower levels of the personality trait than did higher categorical options. Large positive values of β indicated items on which fewer respondents endorsed higher response options, reflecting higher levels of the personality trait, whereas negative values of β indicated items for which more respondents endorsed lower response options. IRT was developed first for application to tests of cognitive ability; the term "difficulty" was directly relevant to this application. It is less relevant to personality items but has become standard for the β parameter.

Table 2
*Parameter Estimates of an Example Item (NEO-FFI Item 2, "I Like to Have a Lot of People Around Me") From the Extraversion Scale of the NEO-FFI Across IQ and Gender Groups*

| | Men | | | | | | Women | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Low IQ | | | High IQ | | | Low IQ | | | High IQ | | |
| Parameters constrained | A | $\beta_1$ | $\beta_2$ | α | $\beta_1$ | $\beta_2$ | α | $\beta_1$ | $\beta_2$ | α | $\beta_1$ | $\beta_2$ |
| None | 0.79 | −2.61 | −0.89 | 1.77 | −0.60 | 2.37 | 0.94 | −2.17 | −0.56 | 1.10 | −1.12 | 1.61 |
| All | 1.20 | −0.67 | 1.48 | | | | 1.00 | −1.07 | 1.02 | | | |
| β | 1.05 | −0.71 | 1.44 | 1.34 | | | 1.23 | −1.03 | 1.08 | 0.82 | | |
| A | 1.23 | −2.57 | −0.71 | | −0.54 | 2.09 | 1.00 | −1.94 | −0.30 | | −1.08 | 1.56 |

| | Low IQ | | | | | | High IQ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Men | | | Women | | | Men | | | Women | | |
| Parameters constrained | A | $\beta_1$ | $\beta_2$ | α | $\beta_1$ | $\beta_2$ | α | $\beta_1$ | $\beta_2$ | α | $\beta_1$ | $\beta_2$ |
| None | 0.79 | −2.14 | −0.42 | 0.94 | −1.17 | 0.44 | 1.76 | −2.35 | 0.62 | 1.10 | −1.12 | 1.61 |
| All | 0.84 | −1.10 | 0.55 | | | | 1.45 | −0.97 | 1.87 | | | |
| β | 0.74 | −0.96 | 0.70 | 1.00 | | | 1.74 | −1.18 | 1.62 | 1.09 | | |
| A | 0.86 | −1.85 | −0.12 | | −1.15 | 0.45 | 1.43 | −1.81 | 0.95 | | −1.18 | 1.74 |

*Note.* NEO-FFI = NEO-Five-Factor Inventory; β = difficulty parameter; α = discrimination parameter; Low = Low-IQ group (range = 70–100); High = High-IQ group (range = 108–123).

There were four models for each item in each grouping. In the first, all of the parameters were freely estimated. In the second, both the difficulty and discrimination parameters were constrained equal. In the third and fourth models, the difficulty parameters, but not the discrimination parameters, were constrained equal, and vice versa. For example, IPIP Item 1, "I am the life of the party," with parameters freely estimated was more discriminating for men with high IQs than for men with low IQs ($\alpha = 1.98$ vs. 1.16). This was not the case for women; the discrimination parameters were 1.52 and 1.58 for high-IQ and low-IQ groups, respectively. This item was more difficult (endorsed less positively) for men with high IQs ($\beta_1 = -2.32$, $\beta_2 = -0.67$, $\beta_3 = 1.95$, $\beta_4 = 6.26$) than for men with low IQs ($\beta_1 = -2.66$, $\beta_2 = -1.91$, $\beta_3 = .14$, $\beta_4 = 2.29$). Similarly, Item 1 was more difficult for women with high IQs ($\beta_1 = -2.24$, $\beta_2 = -.73$, $\beta_3 = 1.96$, $\beta_4 = 4.89$) than for women with low IQs ($\beta_1 = -2.93$, $\beta_2 = -2.13$, $\beta_3 = .65$, $\beta_4 = 3.34$), but the differences were not statistically significant ($p = 1.00$). When the parameters were constrained equal, these differences disappeared, and the parameters took values close to the midpoints among them when freely estimated. (For example, the values for men were $\alpha = 1.43$, $\beta_1 = -1.98$, $\beta_2 = -0.86$, $\beta_3 = 1.36$, $\beta_4 = 4.07$.) When the discrimination but not difficulty parameters were freely estimated, the discrimination parameters were 1.55 and 1.28 for high-IQ and low-IQ groups of men, respectively. For women, discrimination parameters were 1.43 and 1.64 for high-IQ and low-IQ groups, respectively ($p = 1.00$) and the difficulty parameters were all a little more similar to those for the group in which the item was more discriminating ($\beta_1 = -2.10$, $\beta_2 = -.98$, $\beta_3 = 1.24$, $\beta_4 = 3.93$ for men, and $\beta_1 = -2.05$, $\beta_2 = -.92$, $\beta_3 = 1.80$, $\beta_4 = 4.60$ for women). When difficulty but not discrimination parameters were freely estimated, the difficulty parameters were more similar across groups than when they were all estimated freely ($\beta_1 = -2.17$, $\beta_2 = -.61$, $\beta_3 = 1.86$, $\beta_4 = 5.83$ for men with high IQs vs. $\beta_1 = -2.91$, $\beta_2 = -2.12$, $\beta_3 = .03$, $\beta_4 = 2.32$ for men with low IQs, and similarly for women; $p = 1.00$). In the four models, none of tests of parameter constraints showed significant loss of model fit ($p = 1.00$) and, regardless of sex, factor loadings and difficulty parameters appeared to be able to be constrained equal across high- and low-IQ groups.

NEO-FFI Item 2, "I like to have a lot of people around me," with parameters freely estimated was more discriminating and difficult for women with high IQs than women with low IQs and more difficult for women than for men in general; for example, the values for women with high IQs versus low IQs were $\alpha = 1.10$, $\beta_1 = -1.12$, $\beta_2 = 1.61$ versus $\alpha = .94$, $\beta_1 = -2.17$, $\beta_2 = -0.56$, respectively. When parameters were constrained equal again, parameters took values around the midpoints of those when freely estimated ($\alpha = 1.00$, $\beta_1 = -1.07$, $\beta_2 = 1.02$ for women). When difficulty was constrained to be equal, the difficulty parameters were more similar to those for the group in which the item was more discriminating ($\beta_1 = -1.03$, $\beta_2 = 1.08$ for women). When discrimination but not difficulty parameters were constrained to be equal, the difficulty parameters were more similar across groups than when they were all estimated freely ($\beta_1 = -1.94$, $\beta_2 = 0.30$ for women with high IQs and $\beta_1 = -1.08$, $\beta_2 = 1.56$ for women with low IQs). Similar to the IPIP items, none of the tests of parameter constraints from the three constrained models showed significant loss of model fit ($p = 1.00$), and regardless of sex, parameters were able to be constrained equal across groups of people with high and low IQs.

After estimating four different models for each individual item, we needed to know which models were best for them. As mentioned above, we evaluated three model fit indices. Tables 3 and show model fit statistics for the freely estimated and constrained models of IPIP and NEO-FFI items, respectively. The $-2^*$log-likelihood difference statistics were uniformly significant. This is common when sample sizes are large, and this fit statistic gives no recognition to model parsimony. Thus, as planned, we relied primarily on BIC and displayed the $-2^*$log likelihood tests in the table. In general, both the AIC and BIC statistics indicated that both difficulty and discrimination parameters could be constrained equal across groups, but there were some exceptions in which AIC did not indicate that parameters could be constrained equal. This occurred more often in comparisons between men and women than between groups with low and high IQs with sexes combined. BIC gives more weight to model parsimony and tends to recover the correct model more frequently in simulation studies where the correct model is known. This was the reason for giving it more weight in our judgments of model appropriateness, but the inconsistency between results for AIC and BIC does also indicate that, with a larger sample, some differential item functioning across sex might be picked up. This was not the primary focus of this study, but it serves to emphasize the appropriateness of keeping the numbers of men and women equal in our groups with low and high IQs. Tables 3 and 4 show that models with discrimination and difficulty parameters constrained equal across men and women were the best fit for every scale of both the IPIP and NEO-FFI (identified by the minimum value of BIC for that scale, without regard to the others). These findings meant that the IPIP and NEO-FFI items could be considered measurement invariant across groups of people with low and high IQs. Because none of the item parameters showed DIF, we concluded that the IPIP and NEO-FFI items were measurement invariant across groups of people with low and high IQs.

Finding measurement invariance implied valid and meaningful mean differences between people with low and high IQs on all scales of the IPIP and NEO-FFI. Thus, we tested for group mean differences on all scales of both measures (see Table 5) to compare with results from Austin et al. (1997). Consistently, we found statistically significant differences between the mean scores on the Intellect scale of the IPIP and, the Openness to Experience of the NEO-FFI, the Emotional Stability of the IPIP, the Neuroticism (reversed Emotional Stability), and the Agreeableness scales of the NEO-FFI with small to moderate effect sizes. The mean Intellect scores of the IPIP and Openness to Experience scores of the NEO-FFI for people with high IQs were higher than the means for people with low IQs ($p < .001$, with effect size –.41 for the IPIP and –.46 for the NEO-FFI), as was the mean of the Emotional Stability scale of the IPIP ($p < .01$, with effect size –.23), and the mean of the Agreeableness scale of the NEO-FFI ($p < .01$, with effect size –.21). The mean score of the Neuroticism scale of the NEO-FFI for people with high IQs, however, was lower than that for people with low IQs ($p < .001$, effect size .30). We found no significant differences on the other two scales. Moreover, we found that when the sexes were examined separately, there were significant IQ-related differences on various scales of the IPIP and the NEO-FFI. Men with high IQs showed significantly higher

Table 3
*Fit Indices of Models Across Four Groups of Respondents for All Scales of the IPIP*

| Scale | Parameters constrained | Men | | Women | | Low IQ | | High IQ | |
|---|---|---|---|---|---|---|---|---|---|
| | | Log-likelihood | BIC | Log-likelihood | BIC | Log-likelihood | BIC | Log-likelihood | BIC |
| A | None | −3,525.9 | 7,636.9 | −2,864.5 | 6,314.3 | −3,483.7 | 7,552.6 | −2,906.7 | 6,398.5 |
| | All | −3,620.9 | **7,540.2** | −2,921.9 | **6,142.2** | −3,520.5 | **7,339.4** | −2,939.0 | **6,176.3** |
| | β | −3,614.8 | 7,585.4 | −2,913.2 | 6,182.1 | −3,514.2 | 7,384.0 | −2,936.6 | 6,228.9 |
| | A | −3,545.3 | 7,618.3 | −2,871.3 | 6,270.4 | −3,488.1 | 7,504.1 | −2,910.7 | 6,349.3 |
| E | None | −4,127.2 | 8,839.6 | −4,086.3 | 8,757.7 | −4,327.1 | 9,239.3 | −3,886.4 | 8,358.0 |
| | All | −4,219.2 | **8,736.8** | −4,127.6 | **8,553.5** | −4,369.2 | **9,036.8** | −3,926.8 | **8,152.0** |
| | β | −4,211.2 | 8,778.1 | −4,118.2 | 8,592.1 | −4,362.9 | 9,081.5 | −3,919.6 | 8,194.9 |
| | A | −4,139.7 | 8,807.3 | −4,094.3 | 8,716.4 | −4,334.2 | 9,196.2 | −3,890.2 | 8,308.2 |
| ES | None | −3,970.7 | 8,526.6 | −3,992.9 | 8,570.9 | −4,142.1 | 8,869.4 | −3,821.5 | 8,228.1 |
| | All | −4,005.2 | **8,308.7** | −4,013.7 | **8,325.8** | −4,180.8 | **8,660.0** | −3,868.8 | **8,036.0** |
| | β | −4,002.1 | 8,359.9 | −4,011.3 | 8,378.4 | −4,174.8 | 8,705.3 | −3,861.4 | 8,078.5 |
| | A | −3,976.0 | 8,479.8 | −3,995.5 | 8,518.9 | −4,146.7 | 8,821.2 | −3,828.6 | 8,184.9 |
| I | None | −4,167.3 | 8,919.8 | −4,151.8 | 8,888.8 | −4,328.3 | 9,241.7 | −3,990.8 | 8,566.9 |
| | All | −4,249.4 | **8,797.1** | −4,208.7 | **8,715.8** | −4,355.4 | **9,009.1** | −4,019.3 | **8,336.8** |
| | β | −4,243.7 | 8,843.1 | −4,191.2 | 8,738.1 | −4,348.7 | 9,053.2 | −4,015.5 | 8,386.8 |
| | A | −4,174.4 | 8,876.6 | −4,167.7 | 8,863.3 | −4,334.8 | 9,197.5 | −3,996.1 | 8,520.0 |
| C | None | −3,997.2 | 8,579.7 | −3,821.0 | 8,227.2 | −4,022.6 | 8,630.4 | −3,795.6 | 8,176.4 |
| | All | −4,053.8 | **8,406.0** | −3,867.9 | **8,034.2** | −4,054.9 | **8,408.1** | −3,826.4 | **7,951.1** |
| | β | −4,049.6 | 8,454.8 | −3,855.8 | 8,067.4 | −4,048.8 | 8,453.2 | −3,819.8 | 7,995.3 |
| | A | −4,008.7 | 8,545.2 | −3,829.1 | 8,186.0 | −4,032.2 | 8,592.2 | −3,801.8 | 8,131.3 |

*Note.* IPIP = International Personality Item Pool; A = Agreeableness; E = Extraversion; ES = Emotional stability; I = Intellect; C = Conscientiousness; β = difficulty parameter; α = discrimination parameter; BIC = Bayesian Information Criterion; Low = Low-IQ group (range = 70–100); High = High-IQ group (range = 108–123). Fit indices of preferred models are shown in bold.

Agreeableness and Openness to Experience scores of the NEO-FFI than those with low IQs. Men with high IQs also showed higher mean scores on the Intellect and Emotional Stability scales but lower Neuroticism scale scores than those with low IQs. Women with high IQs showed significantly higher mean scores on the Intellect scale of the IPIP and the Openness to Experience scale of the NEO-FFI than those with lower IQs ($p < .001$, with similar effect size –.54).

Table 4
*Fit Indices of Models Across Four Groups of Respondents for All Scales of the NEO-FFI*

| Scale | Parameters constrained | Men | | Women | | Low IQ | | High IQ | |
|---|---|---|---|---|---|---|---|---|---|
| | | Log-likelihood | BIC | Log-likelihood | BIC | Log-likelihood | BIC | Log-likelihood | BIC |
| A | None | −2,821.5 | 5,998.7 | −2,285.5 | 4,926.7 | −2,616.2 | 5,588.1 | −2,490.8 | 5,337.3 |
| | All | −2,854.0 | **5,891.6** | −2,331.9 | **4,847.4** | −2,634.1 | **5,451.9** | −2,502.2 | **5,187.9** |
| | β | −2,837.9 | 5,916.7 | −2,316.7 | 4,874.4 | −2,626.9 | 5,494.8 | −2,496.7 | 5,234.5 |
| | A | −2,828.5 | 5,955.3 | −2,290.9 | 4,880.2 | −2,622.2 | 5,542.8 | −2,496.0 | 5,290.4 |
| E | None | −3,511.8 | 7,448.2 | −3,456.2 | 7,337.0 | −3,504.3 | 7,433.2 | −3,463.7 | 7,351.9 |
| | All | −3,542.2 | **7,302.4** | −3,477.8 | **7,173.7** | −3,554.2 | **7,326.4** | −3,532.7 | **7,283.5** |
| | β | −3,535.6 | 7,358.2 | −3,474.8 | 7,236.4 | −3,544.4 | 7,375.7 | −3,516.0 | 7,318.8 |
| | α | −3,521.6 | 7,399.0 | −3,460.4 | 7,276.6 | −3,511.2 | 7,378.1 | −3,470.2 | 7,296.1 |
| N | None | −2,857.7 | 6,139.9 | −3,263.5 | 6,951.6 | −3,291.4 | 7,007.4 | −2,829.8 | 6,084.1 |
| | All | −2,885.8 | **5,989.7** | −3,281.9 | **6,781.8** | −3,316.7 | **6,851.5** | −2,853.7 | **5,925.4** |
| | β | −2,873.8 | 6,034.5 | −3,277.1 | 6,841.1 | −3,308.6 | 6,904.0 | −2,844.9 | 5,976.6 |
| | α | −2,865.6 | 6,086.9 | −3,271.7 | 6,899.0 | −3,299.4 | 6,954.5 | −2,836.8 | 6,029.4 |
| O | None | −3,546.3 | 7,517.2 | −3,365.3 | 7,155.1 | −3,539.8 | 7,504.1 | −3,371.9 | 7,168.3 |
| | All | −3,584.3 | **7,386.7** | −3,388.0 | **6,994.1** | −3,578.7 | **7,375.5** | −3,396.0 | **7,010.1** |
| | β | −3,572.6 | 7,432.1 | −3,379.7 | 7,046.2 | −3,568.9 | 7,424.7 | −3,391.4 | 7,069.6 |
| | α | −3,557.9 | 7,471.6 | −3,372.8 | 7,101.3 | −3,552.6 | 7,461.0 | −3,378.7 | 7,113.0 |
| C | None | −2,249.5 | 4,889.1 | −2,135.3 | 4,660.8 | −2,134.6 | 4,659.3 | −2,250.2 | 4,890.5 |
| | All | −2,288.0 | **4,776.8** | −2,159.6 | **4,520.1** | −2,159.7 | **4,520.2** | −2,270.9 | **4,742.7** |
| | β | −2,275.0 | 4,813.9 | −2,151.3 | 4,566.6 | −2,147.2 | 4,558.4 | −2,260.7 | 4,785.4 |
| | α | −2,256.6 | 4,840.2 | −2,141.6 | 4,610.3 | −2,139.9 | 4,606.9 | −2,254.6 | 4,836.3 |

*Note.* NEO-FFI = NEO-Five-Factor Inventory; A = Agreeableness; E = Extraversion; N = Neuroticism; O = Openness to experience; C = Conscientiousness; β = difficulty parameter; α = discrimination parameter; BIC = Bayesian Information Criterion; Low = Low-IQ group (range = 70–100); High = High-IQ group (range = 108–123). Fit indices of preferred models are shown in bold.

Table 5
*Means of Low- and High-IQ Groups for the IPIP and NEO-FFI Scales*

| Scale[a] | IQ group | IPIP | | | | | NEO-FFI | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *M* | *SD* | *F* | *p* | Effect size (Cohen's *d*) | *M* | *SD* | *F* | *p* | Effect size (Cohen's *d*) |
| Agreeableness | Low | 41.10 | 5.36 | | | | 45.06 | 5.20 | | | |
| | High | 41.32 | 5.43 | 0.27 | .605 | | 46.19 | 5.39 | 7.04* | 0.008 | −.21 |
| Extraversion | Low | 30.99 | 6.98 | | | | 39.26 | 5.87 | | | |
| | High | 31.58 | 7.10 | 1.10 | .294 | | 38.55 | 5.80 | 2.23 | 0.135 | |
| Emotional Stability (Neuroticism) | Low | 33.82 | 7.69 | | | | 30.08 | 7.23 | | | |
| | High | 35.61 | 7.90 | 8.13* | .005 | −.23 | 27.83 | 7.65 | 14.02** | 0.000 | .30 |
| Intellect/Imagination (Openness to Experience) | Low | 32.86 | 5.20 | | | | 36.78 | 5.34 | | | |
| | High | 35.13 | 5.84 | 25.67** | .000 | −.41 | 39.40 | 6.06 | 32.15** | 0.000 | −.46 |
| Conscientiousness | Low | 38.44 | 5.83 | | | | 47.20 | 5.78 | | | |
| | High | 38.04 | 6.10 | 0.68 | .409 | | 46.28 | 6.10 | 3.65 | 0.057 | |

*Note.* IPIP = International Personality Item Pool; NEO-FFI = NEO-Five-Factor Inventory.
[a] When corresponding IPIP and NEO-FFI scale names differ, NEO-FFI scale name is presented in parentheses.
* $p < .01$. ** $p < .001$ (two-tailed).

## Discussion

All the IPIP and NEO-FFI items functioned similarly in the lower and higher IQ groups. The items also functioned similarly in men and women. Because of this, mean and/or variance differences on the raw scale scores between the groups could be interpreted as real and meaningful and could be compared with each other. When the IPIP and NEO-FFI items displayed measurement invariance, group differences on the items reflected true differences in the five personality scales. Thus, we explored differences in personality scores of the two IQ groups and found that the differences were consistent with previous studies using the NEO-FFI to investigate associations between personality traits and intelligence (Austin et al., 1997; Austin et al., 2002). As in those other studies, we observed that people with high IQs had higher mean Intellect and Openness to Experience and Emotional Stability (lower Neuroticism) scores. This consistency occurred despite our use of different specific measures of the five-factor model's traits. The I and O scales are generally considered to represent traits reflecting intellect or appreciation for culture, which are often associated with cognitive abilities as measured by the MHT. Such associations could confound responses to the I and O items of people with different cognitive abilities if they involved differences in item interpretations that varied with cognitive abilities. Our results, however, indicated that all the scales of the IPIP and NEO-FFI, including the I and O scales, functioned similarly, regardless of the cognitive abilities of respondents. In addition, we also observed associations between sex and personality and found that, for the IPIP, Agreeableness, Emotional Stability, and Intellect scores were higher in both sexes with high IQs. For the NEO-FFI, Agreeableness and Openness to Experience scores were higher in respondents with high IQs, whereas Extraversion, Neuroticism, and Conscientiousness scores were higher in those with low IQs. Our study thus provides additional data attesting to the robustness of these associations and the absence of other associations between intelligence, sex, and the factors of the five-factor model.

Because we found no significant DIF in IPIP and NEO-FFI items, it appeared that those items worked equally well in the low and high IQ groups, and the items were of both comparable difficulty and discrimination in the two IQ groups. Thus, our results did not support the notion that people with lower cognitive ability were less likely to understand item content, to reliably assess their own personalities, and to distinguish between ordered categorical responses of personality scales.

In the process of generating our findings of primary interest, we also made some observations about overall item function that are relevant to the quality of the IPIP's and NEO-FFI's measurements. Were the IPIP and NEO-FFI to be modified in light of these observations, our conclusions about measurement invariance across IQ and sex would have to be revisited. We found extreme patterns of response option endorsements for many items. From our results summarizing response rates, it was clear that people were more likely to use agreement responses (Options 4 and 5) than disagreement responses (Options 1 and 2) in every scale, especially on NEO-FFI items, so that we needed to collapse from five to three response options for the NEO-FFI. For example, for IPIP Item 12r "I insult people," a reversed item from the Agreeableness scale, we found that less than 5% of the respondents endorsed an agreement response (Options 4 or 5), and more than 70% endorsed a disagreement response (Options 1 or 2). Similar results were obtained for NEO-FFI Item 4, an item from its Agreeableness scale: We found that one female with low IQ endorsed disagreement responses (Option 1 or 2), and three men with low and three with high IQ endorsed it. This indicated that people with low and high IQs endorsed these items rarely.

It was notable that the NEO-FFI items performed more poorly than the IPIP items, especially in the low-IQ group. People with low IQ were more likely to answer very positively on NEO-FFI items than on IPIP items. This has been termed acquiescence or social desirability bias in the personality literature. It results in a meaningless increase in scale scores and misinterpretation of testing results. In the present study, the acquiescence bias on several NEO-FFI items among the low-IQ group greatly increased A and C scale scores and reduced the IRT difficulty parameters. The problem was so severe in the NEO for some items that we had to collapse the response options of all NEO-FFI items to retain

sufficient responses in each category. This needs greater attention in future research with this version of the NEO.

Some prior studies have offered explanations for why people may tend to use positive responses. For example, Chen, Lee, and Stevenson (1995) stated that it may be typical of individualists in western countries to view themselves favorably in general, or to wish to portray themselves in favorable lights to others, and people in Asian or collectivist cultures may be less likely to do this. The respondents in all of our groups were Scottish people aged around 70 years at the test dates and, thus, were from a western individualistic culture. Moreover, they were both physically and mentally healthy enough to participate in LBC1947 and, thus, were likely to be aware that they were participating in a study of healthy aging, to be well socialized, and to be patient enough to complete thorough assessments, including especially the IQ test, making it likely that most of them truly were people who rarely, for example, insult others. In addition, all of the scales in the IPIP and all but the Neuroticism scale in the NEO-FFI are coded as positive personal characteristics. Thus, positive response styles were likely to reflect both real situations and positive framing. The reality of these generally positive personality characteristics would be consistent with findings of Srivastava, John, Gosling, and Potter (2003), that people tended to become more agreeable and conscientious with age, and findings of Asendorpf (1998), that there were positive correlations among the Extraversion, Agreeableness, and Conscientiousness scales and between these scales and quality of interpersonal relationships.

Despite the fact that we found no sex-related DIF in any items across the groups, the IPIP instructions make it difficult to interpret sex differences in scale scores and item endorsement frequencies. Respondents were instructed to "Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age." To tell people to compare themselves with people of the same sex and age will bias their judgments relative to the whole population when age and sex effects are present if, in fact, people do apply the suggested reference frame. In addition, as we found no DIF on the NEO-FFI items, our results contradict those of Rammstedt et al. (2010), who found that the five-factor structure was not replicated in people with lower education, unless acquiescence response bias was controlled. In their study, less well-educated people were more likely to endorse options indicating item agreement than were well-educated people. They concluded that measurement was not equivalent across groups of people with different levels of education. Of course, IQ and educational level are not equivalent, but they do tend to be substantially correlated in most samples.

We also observed that people with low IQs tended to use options indicating agreement with items more than did people with high IQs, but the differences were not significant. Moreover, when discrimination and difficulty parameters were freely estimated, the difficulty parameters of options indicating agreement with items (Options 4 and 5) of people with low IQs were lower than those of people with high IQs, but again, the differences were not significant. Thus, although our observations indicated potential differences between the low- and high-IQ groups, we did not find DIF between them. Future studies of DIF on personality scales with larger samples and/or greater differences in IQ might detect such differences. As mentioned above, we grouped low- and high-IQ people by using two ranges of standardized IQ: (70–100) and (108–130). Thus, the difference between the highest IQ of the low-IQ group and the lowest IQ of the high-IQ group was eight points, although the difference in mean IQ between the groups was more than 20. We did specifically exclude respondents having IQs lower than 70 out of concern that they might be unable to complete the personality test properly. Had we retained them in the analysis, they might have revealed DIF. Using samples from different age and cultural groups may also aid future research in providing clearer information about DIF on personality items by IQ differences.

## Conclusions and Recommendations

In prior studies examining associations between personality and cognitive abilities, group differences have been confounded with measurement differences, leaving the effects of cognitive ability on personality scores unclear. Our study used IRT to investigate whether the IPIP and NEO-FFI measured personality consistently in people with differences in cognitive abilities. We found that the IPIP and NEO-FFI items for all five scales were measurement invariant across people with high and low IQs, making it possible to conclude that any differences in IPIP and NEO-FFI scores between people with low and high cognitive abilities reflected true personality trait differences. This should also be explored in demographically different samples and with other personality measures.

## References


Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute, 50,* 277–290.

Asendorpf, J. B. (1998). Personality effects on social relationships. *Journal of Personality and Social Psychology, 74,* 1531–1544. doi:10.1037/0022-3514.74.6.1531

Austin, E. J., Deary, I. J., & Gibson, G. J. (1997). Relationships between ability and personality: Three hypotheses tested. *Intelligence, 25,* 49–70. doi:10.1016/S0160-2896(97)90007-6

Austin, E. J., Deary, I. J., Whiteman, M. C., Fowkes, F. G. R., Pedersen, N. L., Rabbitt, P., . . . McInnes, L. (2002). Relationships between ability and personality: Does intelligence contribute positively to personal and social adjustment? *Personality and Individual Differences, 32,* 1391–1411. doi:10.1016/S0191-8869(01)00129-5

Chen, C., Lee, S.-Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science, 6,* 170–175. doi:10.1111/j.1467-9280.1995. tb00327.x

Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychological Bulletin, 112,* 155–159.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEOFFI)—Professional manual.* Odessa, FL: Psychological Assessment Resources.

Crane, P. K., Gibbons, L. E., Jolley, L., van Belle, G., Selleri, R., Dalmonte, E., & De Ronchi, D. (2006). Differential item functioning related to education and age in the Italian version of the Mini-mental State Examination. *International Psychogeriatrics, 18,* 505–515. doi:10.1017/S1041610205002978

Deary, I. J., Gow, A. J., Taylor, M. D., Corley, J., Brett, C., Wilson, V., . . . Starr, J. M. (2007). The Lothian Birth Cohort 1936: A study to examine influences on cognitive ageing from age 11 to age 70 and beyond. *BMC Geriatrics, 7,* 28.

Deary, I. J., Whalley, L. J., & Starr, J. M. (2009). *A lifetime of intelligence: Follow-up studies of the Scottish Mental Surveys of 1932 and 1947.*

Washington, DC: American Psychological Association. doi:10.1037/11857-000

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29. doi:10.1037/0021-9010.72.1.19

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

Goldberg, L. R. (1999). A broad-bandwith, public-domain, personality inventory measuring the lower-level facets of several Five-Factor models. In I. Mervielde, I. J. Deary, F. de Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, the Netherlands: Tilburg University Press.

Johnson, W., Spinath, F., Krueger, R. F., Angleitner, A., & Riemann, R. (2008). Personality in Germany and Minnesota: An IRT-Based Comparison of MPQ Self-Reports. *Journal of Personality, 76,* 665–706. doi:10.1111/j.1467-6494.2008.00500.x

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Markon, K. E., & Krueger, R. F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior Genetics, 34,* 593–610. doi:10.1007/s10519-004-5587-0

Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus: Statistical analysis with latent variables, user's guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.

Rammstedt, B., Goldberg, L. R., & Borg, I. (2010). The measurement equivalence of Big-Five factor markers for persons with different levels of education. *Journal of Research in Personality, 44,* 53–61. doi:10.1016/j.jrp.2009.10.005

Reise, S. P., Smith, L., & Furr, R. M. (2001). Invariance on the NEO PI-R Neuroticism scale. *Multivariate Behavioral Research, 36,* 83–110. doi:10.1207/S15327906MBR3601_04

Schwarz, G. (1978). Estimating the dimension of a model. *Annual of Statistics, 6,* 461–464. doi:10.1214/aos/1176344136

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the multidimensional personality questionnaire stress reaction scale. *Journal of Personality and Social Psychology, 75,* 1350–1362. doi:10.1037/0022-3514.75.5.1350

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change. *Journal of Personality and Social Psychology, 84,* 1041–1053. doi:10.1037/0022-3514.84.5.1041

Sternberg, R. J., & Ruzgis, P. (1994). *Personality and intelligence.* Cambridge, England: Cambridge University Press.

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *Sage handbook of personality theory and testing. Vol. 2: Personality measurement and assessment* (pp. 254–285). London, England: Sage.

Teresi, J. A., Golden, R. R., Cross, P. S., Gurland, B. J., Kleinman, M., & Wilder, D. E. (1995). Item bias in cognitive screening measures: Comparisons of elderly White, Afro-American, Hispanic and high and low education subgroups. *Journal of Clinical Epidemiology, 48,* 473–483. doi:10.1016/0895-4356(94)00159-N

Toomela, A. (2003). Relationships between personality structure, structure of word meaning, and cognitive ability: A study of cultural mechanisms of personality. *Journal of Personality and Social Psychology, 85,* 723–735. doi:10.1037/0022-3514.85.4.723

Received November 7, 2010
Revision received September 14, 2011
Accepted September 16, 2011 ■