The Suitability of Goldberg's Big Five IPIP Personality Markers in New Zealand: A Dimensionality, Bias, and Criterion Validity Evaluation

Nigel Guenole
Oleksandr S. Chernyshenko
University of Canterbury

This paper presents a series of analyses of New Zealanders' responses to a widely used Big Five personality inventory (Goldberg's IPIP 50 item markers). The suitability of the Big Five Model for the New Zealand work context was investigated by: comparing the fit of the Big Five model with those for other plausible models; establishing measurement equivalence of the Big Five markers across gender groups; and showing evidence of the criterion and construct validity for the five factors with multiple regression analyses and path modeling. Our findings indicated adequate fit for the five-factor model. There was little evidence of measurement bias at the item or scale level. Overall, females scored significantly lower on Emotional Stability and Intellect, but higher on Agreeableness, and Conscientiousness. There were no significant differences between gender groups on Extraversion. The five factors showed great similarity with United States findings in terms of their relation to job satisfaction and contextual performance criteria.

The questionnaire-based individual differences approach to the study of personality searches for a universally applicable set of traits that can explain the inter-individual variation in personality. It is now agreed by many personality researchers that five broad factors account for a large proportion of the variance in self-report personality questionnaires (e.g. Saucier & Goldberg, 2003). These five factors are Neuroticism (Emotional Stability), Extraversion, Openness to Experience (Intellect), Agreeableness, and Conscientiousness. Collectively, they are known as the Five-Factor Model (FFM) of personality or, simply, the Big Five. Because the FFM is found to be robust across cultures, languages, gender and age groups (cf., Hough & Ones, 2001), it provided a common foundation for comparing results from different studies, which, in turn, galvanized personality research around the world. For an informative and brief discussion of the history of the FFM see Goldberg (1993).

The benefits of the FFM in the work context were fully realized only recently. Barrick & Mount (1991) combined the use of the meta-analytic method with the Big Five factor taxonomy to investigate whether personality can predict important organizational outcomes (i.e., job performance or training proficiency). Results indicated that broad personality dimensions are useful in predicting successful performance in many occupational groups. Since then, almost a dozen further meta-analyses investigating personality-job performance links under a FFM framework have been published showing personality is related to job performance, with Conscientiousness and Emotional Stability being the most powerful predictors across occupational groups and performance criteria (Barrick, Mount, & Judge, 2001). The utility of these findings for organizational settings is augmented by the fact that personality variables are not highly correlated with cognitive ability, thus adding incremental validity to selection decisions based on cognitive ability test scores (Schmidt & Hunter, 1998). Moreover, personality-based selection decisions have been found to have less impact against members of ethnic minority groups than cognitive ability-based selection decisions (Day & Silverman, 1989; Gelatly, Paunonen, Meyer, Jackson & Goffin, 1991). As a result, use of Big Five measures in industrial and organizational (I/O) research and practice is now widespread.

While FFM research has gained considerable momentum overseas, New Zealand has generally lagged the United States and Europe in research into the FFM. An exception is Black (2000) who investigated the predictive validity of the five-factor model in a sample of New Zealand Police. In our view, there is a need for research examining the applicability of the FFM in New Zealand, especially in the work context where personnel decisions are routinely based on personality test scores. More specifically, there is a need to test the dimensionality of FFM questionnaires to examine whether the psychometric structure is evident in New Zealander's responses, which, to our knowledge, has not yet been empirically tested. Additionally, the research should address issues of measurement bias, group differences, and predictive validity for selection, so that New Zealand's findings can be confidently compared with overseas research. Such research would be most useful if it were done using an easily accessible and transparent Big Five measure, rather than one of the commercially available instruments for which the content of specific items and scales cannot be publicly disclosed. Therefore, in this paper, we focus our effort on studying Goldberg's (1999) IPIP Big Five 50-item measure. This item set is available at http://ipip.ori.org/ipip/new_home.htm and is a public domain instrument.

Paper Overview

We conducted three studies of Goldberg's IPIP 50-item measure using large samples of New Zealand workers. In the first study, we investigated the dimensionality of Goldberg's Big Five questionnaire by comparing confirmatory factor analysis fit statistics for a range of plausible models. Much of the attractiveness of the five-factor model derives from its ubiquity, and there are, of course, competing models that are potentially equally valid. Our aim was, therefore, to establish whether the Big Five structure of the questionnaire observed overseas was also evident in a population of New Zealand workers.

In the second study, we analyzed measurement equivalence of items across gender groups via mean and covariance structures analysis (MACS). The issue of measurement equivalence was examined because it is becoming a critical psychometric concern for test users. To date, it has been common for cognitive ability and personality testing in New Zealand to occur without comparisons of item functioning across subpopulations, as evidenced by the documentation of test publishers. In this study, only responses of males and females were compared, as small sample sizes for Maori and Pacific Island Peoples precluded analysis of ethnic groups. We note that ethnic group measurement equivalence is an important research area in the New Zealand context, and that procedures demonstrated in this article for gender are also applicable for ethnic measurement equivalence studies.

Our third study targeted criterion validity evidence for Goldberg's measure. A key reason for the advancement of the FFM in applied settings has been the consistency of findings that it yields regarding predictions of various work outcomes. In our study, we assessed how the five factors related to the two most important dependent variables in industrial/organizational psychology research, job satisfaction and job performance. The criterion validities of Big Five personality scales were evaluated using multiple regression analyses as well as a more sophisticated path modeling methodology.

Study 1: Dimensionality of Goldberg's IPIP Questionnaire

One of Goldberg's goals for the factor markers was to provide a parsimonious set of items that would generate the five-factor target structure, against which alternative theoretical positions and other personality questionnaires could be compared. However, if the proposed theoretical structure were not evident in the responses to items, score interpretations for any purpose (i.e., research, selection, or development) would be dubious. The aim of this analysis was, therefore, to establish whether the five factor structure evident in overseas responses to the factor markers was also evident in New Zealand collected responses. We expected to observe this structure, given that it has been replicated across many cultures (Hough & Ones, 2001).

The most appropriate methodology for verifying the underlying structure of an inventory is confirmatory factor analysis (CFA). It is theory driven and allows researchers to confirm or disconfirm a priori models. Although some personality researchers criticize CFA as too stringent and argue in favor of exploratory factor analysis (e.g., Borkenau & Ostendorf, 1991; McCrae, Zonderman, Costa, Bond & Paunonen, 1996), the CFA method, when implemented appropriately, provides the most powerful factor structure evidence. An important step in the CFA analysis is not only to confirm the hypothesized structure, but also to disconfirm alternative plausible structures. The size of the model-data fit indices governs such confirmation decisions. In the case of Goldberg's IPIP measure, the theorized structure is, of course, the Five Factor model. However, a number of alternative models are also plausible. Thus, we explicitly test the fit of competing models to our data and compare it to the fit of the Big Five model.

Goldberg's Five-Factor Model

The factors in Goldberg's model are Extraversion, Agreeableness, Intellect, Conscientiousness, and Emotional Stability. This model was expected to show the strongest fit to the data. We note here that there are two widely recognized Five Factor models, that of Goldberg (1990) and that of McCrae & Costa (1987). These models are comparable, the minor differences being in the naming conventions, (e.g. Emotional Stability and Intellect factors in Goldberg's model are called Neuroticism and Openness to experience in McCrae & Costa's model), and the claimed theoretical bases of the models.

Other Models

We tested the fit of the Goldberg model against the fit of two more parsimonious models loosely based on other plausible theoretical models. The first alternative model was the Eysenck's (1991, 1992) model. Eysenck has perhaps been the primary personality theorist opposing the FFM. The Eysenck model is a three-factor model in which Neuroticism and Extraversion scales measure their namesakes in the Big Five model, while Agreeableness and Conscientiousness constitute the third factor, Psychotocism. The remaining factor of the Big Five, Intellect, is not well accounted for by Eysenck's model (Hough & Ones, 2001). Consequently, for our purposes, we formed a variation of Eysenck's model and tested the following four-factor model: Extraversion, Neuroticism, Psychotocism (Conscientiousness and Agreeableness items loading on the same factor), and Intellect.

The second model, which we call "Integrity model," is loosely based on the work of Ones and Viswesvaran (2001) who noted that Integrity, as measured by overt Integrity tests, correlated highly with Conscientiousness, Agreeableness, and Emotional stability. To test this alternative, and more parsimonious model, we analyzed the following two-factor model that we call the Integrity model: Conscientiousness, Agreeableness, and Emotional Stability loading on the first factor, and Extraversion and Intellect on the second factor. We note here that in another paper, Hough and Ones (2001) suggested that Agreeableness, Emotional Stability and Conscientiousness factors can be considered a compound trait they called "customer service orientation", the difference from Integrity being the relative contributions of the factors to the compound trait. Our test model is also similar to one proposed by Digman (1997) who described the constellation of Agreeableness, Emotional Stability and Conscientiousness as type Alpha personality.

Method

Participants - A total of 452 individuals participated in this part of the study. The majority (229) of the participants came from two call center work environments in finance and healthcare. The remainder came from both management and non-management level jobs across a variety of business organizations that agreed to participate in the study. All participants were employed at the time of the research. Overall, a wide range of job levels was represented in this sample, including management level jobs, and participants' job and personal characteristics were similar across call-center and non-call center groups.

There were 251 females (mean age =36.06 years, standard deviation = 9.81 years) and 201 males (mean age = 38.98 years, standard deviation = 9.69 years). The ethnic composition of the sample was as follows: Asian (21 respondents); European (384 respondents); Indian (11 Respondents); Maori (25 Respondents); Pacific Island (10 respondents), and Other (1 respondent). Of the 452 participants, 286 reported possessing some form of tertiary qualification.

All participants were recruited via e-mail. The data were collected over the Internet as a part of a larger survey development project. Subjects were allowed to complete the questionnaire in their own time, and they were permitted to log in more than once if completion of the questionnaire required more than one Internet session. They were not permitted to alter answers once they submitted responses.

Measures - The Big Five dimensions of personality were assessed using Goldberg's IPIP 50-item measure (Goldberg, 1999). The measure is comprised of short sentences describing various behaviors associated with each of the Big Five dimensions (i.e., Extraversion, Conscientiousness, Agreeableness, Emotional Stability, and Intellect). Each Big Five scale contains 10 items paired with a 5-point Likert response scale (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, and 5=strongly agree). The reliabilities of the Big Five scales ranged from 0.78 to 0.88. All relevant item and scale statistics (i.e., item means, standard deviations, and item-total correlations) are available upon request from the authors.

Analyses - LISREL 8.54 was used to fit the Goldberg five-factor model, the variation of Eysenck's model, and the Integrity model to the dataset. We estimated the models using weighted least squares estimation with asymptotic and polyserial covariance matrices, which is appropriate when the responses are ordinal and multivariate normality is questionable (Byrne, 1998). We ran confirmatory factor analyses constraining factor loadings as follows.

Goldberg's structure required a five-factor model constraining loadings on all factors to zero other than for the parent factor for the item, on which the loading was set to one. The Eysenck variation required fitting a four-factor model in which the factors were formed by (1) Extraversion, (2) Agreeableness and Conscientiousness, (3) Emotional Stability; and (4) Openness. Again, items were set to one on the parent factor and zero on all other factors. Finally, the Integrity model was tested with a two-factor CFA model in which items measuring Conscientiousness, Agreeableness, and Emotional Stability were set to one on the first factor and zero on the second factor, while items measuring Intellect and Extraversion were set to one on the second factor but zero on the first.

For each of the three models, we interpreted the fit in light of the issues associated with item level structural equation modeling – in particular, the poor fit often evident when item level data are used. We report multiple indices in addition to the model χ^2 because its sensitivity to sample size can lead to rejection of theoretically appropriate models (e.g. Byrne, 1998). Additional indices considered included the expected cross validation index (ECVI); the root mean square error of approximation (RMSEA; Steiger & Lind, 1980); Jöreskog & Sörbom's (1986) Standardized Root Mean Square Residual (SRMSR); Bentler's (1990) Comparative Fit Index (CFI); and the Goodness of Fit index (GFI; Tanaka & Huba, 1984). Because our models are not nested, ECVI was used to assess the likelihood that the model in question would cross-validate on equivalent sized samples from the population. According to Byrne (1998), models having the smallest ECVI values exhibit the greatest potential for replication.

An important issue regarding the stability of the CFA results is whether to factor analyze individual items or multi-item composites (parcels). Many researchers (e.g., Bernstein & Teng, 1989; Catell, 1947) prefer to use parcels, because they are more reliable than individual items and better resemble continuous data assumed by CFA. Thus, we proceeded to examine the fit of the three structures described above using parcels. For each of Goldberg's five IPIP scales, we created three-item parcels (unit weighted sums), one having four items and the other two having three items each. Parcels were created by allocating items one to four to parcel one, items five to seven to parcel two, and the final three to parcel three. The CFA loading patterns indicated that the ordering of the scale items had no statistical meaning, and therefore, this parceling strategy is essentially random.

Results

Fit statistics for the three models examined are presented in Table 1. Overall, the five-factor model (Goldberg model) showed comparatively better fit than its alternatives. Specifically, the ECVIs for the Goldberg model were 8.06 (item level) and .09 (parcel level) as compared to 9.05 and 1.41 for the Eysenck model, and 9.82 and .14 for the Integrity model, thus indicating that the five-factor model was most likely to replicate in future research having equivalent size samples. Furthermore, fit statistics for Goldberg's model at the parcel level were all near or above recommended critical values representing excellent fit (i.e., GFI = .97, SRMSR = .09, NNFI = .94, and CFI = .96).

At the item level, fit for all models was less adequate (e.g., for the Goldberg model the CFI was .81, the NNFI was .8, and the SRMSR was .21). This was not particularly surprising, given previous research findings regarding the difficulty of conducting CFA using item level data (e.g., Borkenau & Ostendorf, 1991; McCrae, Zonderman, Costa, Bond & Paunonen, 1996). It appears that having only five discrete response options instead of the more continuously distributed item scores provided by parceling adversely affected the observed CFA-based fit statistics. Note that parceling mitigated this effect.

Discussion

Our results indicated that the five-factor structure (i.e. the Goldberg structure) exhibited the best model fit for the 50-item Goldberg's IPIP questionnaire. Alternative structures, although plausible and more parsimonious, showed worse fit as evidenced by ECVI and other fit indices. Forming item parcels has greatly improved fit, suggesting that researchers interested in conducting CFA analyses using Goldberg's IPIP questionnaire may want to prefer that option in the future. Overall, it appears that the dimensionality of the 50-item IPIP questionnaire was consistent across New Zealand as it is in the United States, namely that five factors were needed to account for the observed variability in item responses.

Table 1. CFA Fit Statistics for Structure Models

Level of analysis	Model	χ^2	df	RMSEA	ECVI	SRMSR	NNFI	CFI	GFI
		Fit Statistics							
Item	Goldberg	3413.42	1165	.07	8.06	.21	.80	.81	.90
	Eysenck	3871.35	1169	.07	9.05	.26	.76	.77	.89
	Integrity	4222.56	1172	.08	9.82	.31	.73	.74	.88
Parcel	Goldberg	365.66	80	.09	.99	.09	.94	.96	.97
	Eysenck	564.30	84	.11	1.41	.17	.91	.93	.96
	Integrity	847.70	87	.14	2.03	.25	.86	.89	.94

Note: N = 452; Item = analyses were conducted using 50 item response data; Parcel = analyses were conducted using 15 parcels (3 for each scale); = chi-square statistic; df = degrees of freedom; RMSEA = root mean square error of approximation; ECVI = expected cross validation index; SRMSR = standardized root mean square residual; CFI = comparative fit index; and GFI = goodness of fit index.

Note that these 50-item markers were intended to be general measures of the Big Five factors, perhaps suitable for research purposes or initial selection screening only. This questionnaire was not designed to provide in-depth personality assessment for personnel selection or feedback purposes. For applications such as these, one should use longer and more reliable scales, or, perhaps, measures designed to assess narrow (facet level) personality traits; examples of these are available on the International Personality Item Pool website.

Study 2: Measurement Equivalence of Goldberg's IPIP Questionnaire: Gender Group Comparisons

Examination of group differences is only meaningful if there is measurement equivalence between groups. Measurement equivalence is satisfied when individuals with equal standing on the trait measured by a test, but sampled from different subpopulations, have the same expected test scores (Drasgow, 1987). Lack of measurement equivalence is often referred to as measurement bias. Presence of bias is undesirable, because, by definition, members of one group will be favoured over members of another group in terms of their item or scale scores even though they have equivalent standing on the underlying trait. In addition, bias can contribute to the observed mean score group differences, confounding comparisons across groups (Stark, Chernyshenko, & Drasgow, 2004).

At present, detecting measurement bias can be done with two classes of methods, those based on item response theory (IRT) and those based on confirmatory factor analysis (for a detailed discussion, see Raju, Laffitte, & Byrne, 2002). Because IRT-based analyses require larger sample sizes than were available to us, a CFA-based bias detection procedure, mean and covariance structures analysis (MACS; Sörbom, 1974) was chosen for this investigation. We conducted our analyses only for gender groups only because ethnic composition of the sample did not provide ethnic groups of sufficient size.

Costa, Teracciano & McCrae (2001) noted that there have been two distinct theoretical positions advanced for gender differences in personality. The first is biological, and suggests that there are distal causes of the observed differences that have been shaped by natural selection. The second theory advances more proximal social psychological explanations for the observed differences, for example, societal role models. Regardless of cause, the practical implications of personality score differences are that there will be *impact* in any situation in which employment selection decisions are made solely on the basis of personality questionnaire scores. For example, if females were found to score higher on the Conscientiousness factor, then decisions made based on Conscientiousness scores would favour females. However, engaging in such comparisons at the construct level is meaningless, unless measurement equivalence of scales across groups has been established a priori. Consequently, we examined the measurement equivalence of the individual scales of Goldberg's IPIP questionnaire in an effort to report bias-free, gender scale score differences and evaluate the extent of the expected impact. We were unable to locate other studies that investigated differential item functioning on Goldberg's 50-item IPIP measure.

Method

Participants - The participants in this analysis were the same group of participants as study one.

Analyses - In MACS analysis, the single common factor model can be written as:

$$x_{ij} = \tau_i + \lambda_i \xi_j + \varepsilon_{ij} \qquad (1)$$

where

x represents a score on an item

i represents items, indexed i - 1,2,...,10;

j represents respondents;

τ represents a vector of item intercepts;

ξ_j represents the factor score for respondent j;

λ_i represents the loading of item i on the common factor ξ ;

ε_{ij} is the unique factor score (often referred as error) for respondence j on item i;

x_{ij} represents a respondent's item score.

Essentially, if there were no measurement bias across gender groups, then all τ_i and λ_i for males would be equal to those for females. To test for the equivalence of loadings and intercepts across groups, one must specify a baseline model where all parameters are free to vary except a *reference item* (i.e., *referent*) whose loading is set equal to 1 in both groups. In addition, it is necessary to constrain the intercepts for the referent to be equal across groups and the latent mean for one group to be zero. The inclusion of these constraints is needed for identification and linking purposes (see Byrne, 1998; Jöreskog & Sörbom, 1996; Reise, Widaman, & Pugh, 1993). The next step is to specify a series of compact (constrained) models where, in each case, the respective loadings and intercepts for one item at a time (the studied item) are constrained equal across groups. The baseline and constrained models are then estimated in succession to obtain a chi-square goodness of fit statistic for each. Because each constrained model is nested within the baseline model, the difference in the normally weighted chi-square between two models is itself distributed as a chi-square statistic with two degrees of freedom (two extra parameters are estimated in the baseline). If a statistically significant chi-square difference is observed for a given baseline and constrained model comparison, the hypothesis of equivalence of the constrained parameters is rejected and the studied item is viewed as showing measurement bias.

The MACS bias detection procedure described above was implemented using LISREL 8.54. Biased items were identified and removed from the IPIP questionnaire prior to scale score comparisons. Next, bias-free mean scale scores were computed for each gender group and the female score was subtracted from the male score to estimate impact. For each scale, effect sizes were also calculated to provide scale independent estimates.

Results

Table 2 presents MACS bias detection results for Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect scales of Goldberg's IPIP questionnaire. In column 3 of each table, we report the chi-square statistics for the baseline model (referent item is shown in parentheses) and for the nine constrained models, which had the studied item constrained equal across gender groups. Columns 4 and 5 show the chi-square differences between the baseline and each of the constrained models, and the associated p-values. If the observed p-value was smaller than the critical p-value (0.05), the studied item was flagged as biased.

Our results indicated that 50-item markers showed relatively little measurement bias across groups of male and female New Zealand workers. No biased items were found for Agreeableness and Conscientiousness. Bias was observed for just one item on the Intellect scale (item 2), two items on Emotional Stability (Items 3 and 5) and three items on the Extraversion scale (item 2, 8 and 9). Post-hoc analyses indicated that in these instances, bias occurred mainly on the intercept parameters after the two groups were placed on a common metric. In a selection situation, the impact of any bias would be small given that so few of the items were biased.

Note that the presence of biased items, although few, indicated that male-female total scale score differences could not be readily compared for Extraversion, Emotional Stability and Intellect scales, unless biased items were excluded. Table 3 reports bias-free scale score differences and their respective effect sizes with respect to females. In addition, we computed effect size confidence intervals using the approach described by Hedges & Olkin (1985) to see whether the observed impact was significant. Results indicated that females scored significantly lower on the Emotional Stability and Intellect scales, but significantly higher on the Agreeableness and Conscientiousness. There was no difference in the Extraversion scores. These results were similar to those reported for samples in the United States and Europe, where males were found to score higher on the Emotional Stability and Intellect (Openness) factors, and lower on Agreeableness (Caprara, Caprara, & Steca, 2003; Feingold, 1994). However, contrary to previous research, males in our study did not score higher on the Extraversion factor.

Discussion

In sum, Goldberg's IPIP questionnaire seemed to function similarly across groups of New Zealand males and females. Bias was observed for only 6 out of 50 items. This indicates that the instrument can be used reasonably confidently in either subpopulation. While we observed significant total score differences on four of the five Big Five scales, they reflected the actual differences in latent distributions and were not caused by measurement bias. We note that in a compensatory selection situation, where a single composite score is formed by summing scores across personality and other predictor constructs, the observed gender differences may cancel out if there are gender differences in opposite directions on the alternative predictors.

Table 2. Measurement Equivalence Results for the Five Goldberg's Scales

Extroversion	Scale/Item Content	χ^2	$\Delta \chi^2$	p-value
Base (4)	Keep in the background.	338.5278		
Item 1	Am the life of the party.	340.9109	2.3831	.3038
Item 2	**Don't talk a lot.**	**345.3515**	**6.8237**	**.0330**
Item 3	Feel comfortable around people.	341.2014	2.6736	.2627
Item 5	Start conversations.	340.0209	1.4931	.4740
Item 6	Have little to say.	344.0421	5.5143	.0635
Item 7	Talk to a lot of different people at parties.	339.8266	1.2988	.5224
Item 8	**Don't like to draw attention to myself.**	**349.3241**	**10.7963**	**.0045**
Item 9	**Don't mind being the center of attention.**	**344.7418**	**6.2140**	**.0447**
Item 10	Am quiet around strangers.	341.8062	3.2784	.1941
Agreeableness	Scale/Item Content	χ^2	$\Delta \chi^2$	p-value
Base (4)	Sympathize with others' feelings.	258.4855		
Item 1	Feel little concern for others.	259.1964	.7109	.7009
Item 2	Am interested in people.	261.8348	3.3493	.1874
Item 3	Insult people.	258.5333	.0478	.9764
Item 5	Am not interested in other people's problems.	260.7644	2.2789	.3200
Item 6	Am not really interested in others.	260.1015	1.6160	.4457
Item 7	Have a soft heart.	258.7110	.2255	.8934
Item 8	Take time out for others.	259.9155	1.4300	.4892
Item 9	Feel others' emotions.	258.7025	.2170	.8972
Item 10	Make people feel at ease.	263.1873	4.7018	.0953
Conscientiousness	Scale/Item Content	χ^2	$\Delta \chi^2$	p-value
Base (7)	Like order.	353.3600		
Item 1	Am always prepared.	356.1139	2.7539	.2523
Item 2	Leave my belongings around.	353.3996	.0396	.9804
Item 3	Pay attention to details.	358.0730	4.7130	.0948
Item 4	Make a mess of things.	353.7343	.3743	.8293
Item 5	Get chores done right away.	354.5368	1.1768	.5552
Item 6	Often forget to put things back in their proper place.	354.5646	1.2046	.5476
Item 8	Shirk my duties.	354.5610	1.2010	.5485
Item 9	Follow a schedule.	354.6293	1.2693	.5301
Item 10	Am exacting in my work.	354.1204	.7604	.6837
Emotional Stability	Scale/Item Content	χ^2	$\Delta \chi^2$	P-value
Base (8)	Have frequent mood swings.	244.8430		
Item 1	Get stressed out easily.	246.7220	1.8790	.3908
Item 2	Am relaxed most of the time.	247.6758	2.8328	.2426
Item 3	**Worry about things.**	**254.9837**	**10.1407**	**.0063**
Item 4	Seldom feel blue.	245.3991	.5561	.7573
Item 5	**Am easily disturbed.**	**254.5410**	**9.6980**	**.0078**
Item 6	Get upset easily.	250.5085	5.6655	.0589
Item 7	Change my mood a lot.	245.2293	.3863	.8244
Item 9	Get irritated easily.	246.8320	1.9890	.3699
Item 10	Often feel blue.	246.5667	1.7237	.4224
Openness	Scale/Item Content	χ^2	$\Delta \chi^2$	P-value
Base (5)	Have excellent ideas.	472.8602		
Item 1	Have a rich vocabulary.	473.3206	.4604	.7944
Item 2	**Have difficulty understanding abstract ideas.**	**483.5458**	**10.6856**	**.0048**
Item 3	Have a vivid imagination.	475.1445	2.2843	.3191
Item 4	Am not interested in abstract ideas.	476.8514	3.9912	.1359
Item 6	Do not have a good imagination.	475.7118	2.8516	.2403
Item 7	Am quick to understand things.	474.8191	1.9589	.3755
Item 8	Use difficult words.	475.4113	2.5511	.2793
Item 9	Spend time reflecting on things.	476.6910	3.8308	.1473
Item 10	Am full of ideas.	473.2906	.4304	.8064

Note: Entries in bold indicate biased items

Study 3: Criterion Validity of the Goldberg's IPIP Scales

Because the utility of the Five Factor model stems in large part from its ability to predict a wide range of work outcomes and demographic criteria, it makes sense, when evaluating a Big Five questionnaire for New Zealand, to examine the relationship of its scales with dependent variables. This research is particularly important given the prevalence of personality testing for selection in New Zealand, and the relative shortage of New Zealand based research on predictive validity of personality scales (the majority of commercially available instruments in NZ were validated in the UK or USA).

In this study, we conducted several multiple regression analyses in which Big Five markers were used to predict three dependent variables frequently studied in I/O psychology: *job satisfaction, counter-productivity,* and *organizational citizenship behaviour.* Multiple regression analyses rather than simple correlations are reported because they better reflect the nature of a typical selection process where multiple factors are used simultaneously to make judgments about candidates.

Additionally, two of the five factors, Extraversion and Emotional stability, were embedded within a path model of antecedents and consequences of job satisfaction in order to evaluate their relationships with multiple criteria. Path modeling is a powerful multivariate technique that allows specification of a series of ostensibly causal relations within a single model and the testing of the entire system of variables to enable a clearer conceptualization of theory under study (Byrne, 1998). We describe the path model in more detail below.

Path Model of Antecedents and Consequences of Job Satisfaction.

Credé, Chernyshenko, Stark, Dalal, and Bashshur (2003) proposed a general causal model for work attitudes in an attempt to integrate various classes of antecedents and consequences of job satisfaction, studied previously in relative isolation (see Figure 1 for a graphical representation of their model). Relevant to our study are two classes of antecedents: trait dispositions and affective workplace events. Dispositions are viewed as enduring, stable personal characteristics, capable of exerting generalized effect on one's attitudes (likes and dislikes). Judge, Bono,

Table 3. Bias-Free Gender Difference Statistics

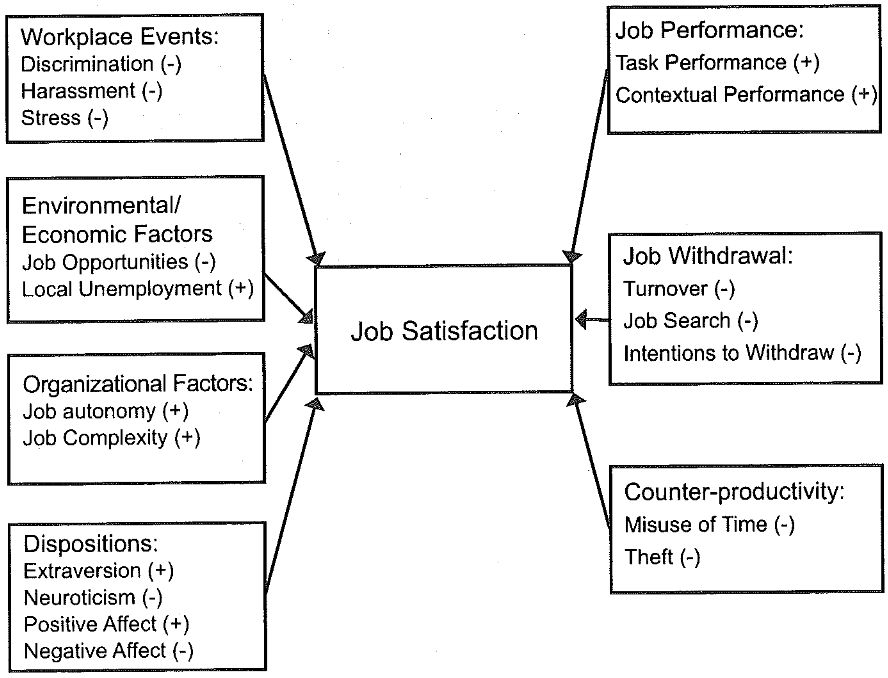
Factor	Female		Male		Effect Size		95% CI	
	Mean	*Std dev*	*Mean*	*Std dev*	*d*	*Std error*	*Lower*	*Upper*
Extroversion	24.71	5.44	23.93	5.74	.14	.10	-.05	.33
Agreeableness	41.80	4.84	39.18	5.32	.52	.10	.33	.71
Conscientiousness	38.23	6.28	36.63	6.11	.26	.10	.07	.44
Emotional Stability	28.17	6.54	29.54	6.16	-.21	.10	-.40	-.03
Openness	34.61	4.85	36.15	4.91	-.32	.10	-.50	-.13

Note: * = scale score comparisons for Extraversion, Emotional Stability and Intellect are based on reduced items due to observed bias; all effects reported in relation to females.

and Locke (2000) found that most relevant dispositions for predicting job satisfaction were Extraversion and Neuroticism, which in our study are represented by Extraversion and Emotional Stability scales of Goldberg's IPIP questionnaire. Job satisfactions are generally defined as a set of affective responses to job characteristics (Hulin & Judge, 2003). As extroverted individuals are known to experience feelings such as cheerfulness and positive emotions, while low emotional stability individuals are known to experience symptoms such as anxiety and depression, the hypothesized link with job satisfaction is theoretically sensible. Based on the work of Credé et al (2003), and Hulin & Judge (2003) both scales were expected to have a positive relationship with job satisfaction.

Affective workplace events are events that individual employees may experience, but that are not an inherent component of the job, or part of the standard job description. Although a multitude of events have been found to influence job attitudes, in our study, we focused on stress only (events such as harassment and discrimination are difficult to study using random samples because the base rate of these events is usually low). Stress was assessed by a measure of overall work stress (Stress-in-General scale, Stanton, Balzer, Smith, Parra, & Ironson, 2001) which was anticipated to have negative relations with job satisfaction. In addition, because Emotional Stability was likely to shape one's sensitivity to stressful events, it was hypothesized to have a direct negative path to stress.

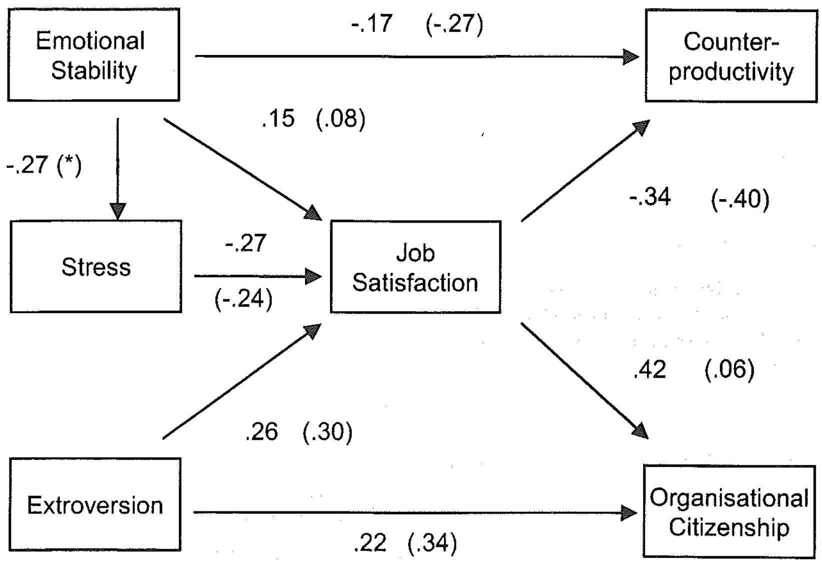
Figure 1. Credé *et al.*'s (2003) general model of antecedents and consequences of job satisfaction.



Two outcome variables were included in the model, organizational citizenship and counter-productivity. Organizational citizenship includes behaviors such as helping coworkers, volunteering, and speaking highly of one's organization. As well as being linked with organizational performance (Harter, Schmidt, & Hayes, 2002), these behaviors are considered intrinsically desirable, and contextual performance is rapidly becoming one of the most studied criteria in I/O psychology. Studies have found that organizational citizenship contributes to overall job performance ratings (Motowidlo & Van Scotter, 1994) and is likely to be influenced by positive dispositions and attitudes rather than job knowledge (Borman & Motowidlo, 1997). Consequently, in our study, we hypothesized direct paths from job satisfaction and Extraversion to organizational citizenship.

Our second outcome variable, counter-productivity (including theft, misuse of time, alcohol or drug use, and insubordination), is also likely to be caused by dispositions and job attitudes, but the path coefficients are expected to be negative. Note also that previous research by Crede *et al.* (2003) found Emotional Stability to be a key dispositional variable predicting counter-productivity. Thus, for our model, we posited direct paths from emotional stability and job satisfaction to counter-productivity. Note that in the model, we specified both direct and indirect links (through job satisfaction) between personality traits (Extraversion and Emotional Stability) and outcomes. If direct links were found to be insignificant, then job satisfaction acted as a mediator. The resulting path model is a smaller version of the model

Figure 2. New Zealand path model results



Note: Crede *et al*'s results are in parentheses. Neuroticism is Emotional Stability in Goldberg's model; all coefficients reported were significant; Emotional Stability to General Stress was not examined by Crede *et al.*

of Credé *at al.* (2003), and it is depicted graphically in Figure 2.

Method

Participants - We collected data on these additional measures for a subset of the original 452 subjects. The subset consisted of employees at the two call centers, which agreed to participate in this phase of the research. There were 229 participants in total, 135 females (mean age =34.87 years, standard deviation = 9.48 years) and 94 males (mean age = 40.50 years, standard deviation = 9.77 years). The ethnic composition of the sample was as follows: Asian (10 respondents); European (194 respondents); Indian (4 Respondents); Maori (15 Respondents), and Pacific Island (6 respondents). English was a second language for 11 of the 229 participants. Of the 229 participants, 120 reported possessing some form of tertiary qualification. The conditions for data collection were the same as for study one.

Measures

Job Satisfaction - A 10-item version of the Satisfaction with Work subscale of the Illinois Job Satisfaction Index (Chernyshenko et al., 2003) was used to measure job satisfaction. It was designed to measure both affective and evaluative components of attitudes toward one's current job. For each item, respondents are asked to indicate the degree to which an item accurately describes their work situation. A four-point scale ("Strongly Disagree"; "Disagree"; "Agree"; "Strongly Agree") is used. Internal consistency reliability was .86.

Emotional Stability and Extraversion - 10-item Big Five marker scales from the International Personality Item Pool (Goldberg, 1999) were used to measure Emotional Stability and Extraversion personality factors. Responses were collected using a 5-point scale. Internal consistency reliability was .87 for Emotional Stability and .86 for Extraversion.

Stress - General work stress was assessed with the 15-item Stress in General scale (SIG; Stanton, Balzer, Smith, Parra, & Ironson, 2001) that uses a three-point scale ("Yes", "?", "No"). In this study, the internal consistency reliability estimate was .86.

Organizational Citizenship Behaviors - A 12-item OCB measure (adopted from work by Borman and Motowidlo, 1997) was used to assess a variety of important behaviours that are generally not specified in job descriptions, but are important for successful functioning of an organization. Respondents were presented with a list of OCB behaviors, such as helping co-workers, volunteering, and speaking highly of one's organization, and were asked to indicate how often they engaged in these behaviors. A five-point Likert format was used for all items from 0 (Never) to 4 (Many times). The internal consistency reliability estimate for the OCB scale in this study was .83.

Counter-productivity - A 10-item Work Withdrawal scale developed by Hanisch and Hulin (1991) was used as measure of counter-productivity. It measures absenteeism, tardiness, and other behaviors reflecting employee desires to avoid work tasks and the work environment. Items described specific work withdrawal behaviors and the respondents were asked how often they have engaged in these behaviors on a 5-point scale from 0 (Never) to 4 (Many times); internal consistency reliability of the measure in this study was .77.

Analyses

SPSS 11.0 was used to regress personality scale scores on three dependent variables – job satisfaction, counter-productivity, and organizational citizenship behaviour. For the path analysis, LISREL 8.54 was used. Following the recommendations of Anderson and Gerbing (1988), measurement models for each of the six variables in the path model were fitted first. The term "measurement model" refers to a class of CFA models concerned with evaluating how well a specific latent variable is represented by its indicators (items). If the fit is good, it is assumed the variable is measured with negligible error and a simple sum of individual items (i.e., scale total score) can be used for subsequent path analyses.

Results

Table 4 presents multiple regression results for Goldberg's Big Five IPIP markers predicting job satisfaction, counter-productivity, and organizational citizenship behaviors. As evident by high multiple correlation coefficients, Big Five scales showed high criterion validities. The highest multiple-correlation was with citizenship behaviors (R = 0.50), followed by job satisfaction and counter-productivity (R = 0.40 and R = 0.31, respectively). In particular, Extraversion and Emotional Stability were strong predictors of job satisfaction; Emotional Stability was a strong predictor of work withdrawal; and Extraversion and Agreeableness were strong predictors of organizational citizenship behaviour. Importantly, in all three analyses, having all five scales in the regression equation improved the overall prediction of work outcomes. While these results support the use of the five factors for prediction of performance outcomes, as pointed out by one reviewer, regression often overestimates the relation between predictor and criterion variables by over-fitting the model to unique aspects of the dataset.

Fit statistics for the six measurement models are presented in Table 5. As can be seen from the statistics that all measurement models showed good fit to data (e.g., all GFI values were above .90). The appropriateness of representing all variables with their respective scale scores was further augmented by an excellent fit found for our path model. The x^2 statistic was insignificant (4.83 with 8 degrees of freedom), RMSEA was near zero, and the GFI was .99.

Resulting coefficients for the path analysis are shown in Figure 2. For comparison, in parentheses, we reported path coefficients from the Credé et al. study of more than 1000 US workers, except for the Emotional Stability – stress path, which was not studied there. Overall, the direction and the strength the relationship between variables was remarkably similar to US research. Focusing on personality variables, it can be seen that the magnitude of the relationship between Extraversion and job satisfaction was .26 in the current research and .30 for the Credé et al's study. Paths between Extraversion and contextual performance were .22 and .34, respectively. Emotional Stability was also found to have predicted patterns of relationships with outcomes, as evident by the .15 path with job satisfaction and the -.17 path with counter-productivity; in the US context, these paths were equal to .08 and -.27. The only departure between two countries was in the strength of relationship between job satisfaction and citizenship (.42 vs. .06), but, although intriguing, was not relevant to our personality investigation.

Table 4. Regression Analyses of the Big Five Marker Scales on Criterion Measures

	Big Five Factor					
Criterion Measures	EXTRO	AGREE	CONS	EMOT	OPEN	R1
Job Satisfaction	.22*	.13	.07	.20*	. .02	.40
Counter Productive Work Behaviour	-.08	-.04	-.13	-.24*	.11	.31
Organizational Citizenship Behaviour	.26*	.27*	.19	.01	.03	.50

Note: N= 229; Table contains standardized beta weights when all five factors are entered in regression equation; OCBs = organizational citizenship behaviors; R1 = multiple correlation between the five factors and each of the criterion variables; * indicates a significant predictor at 0.05 level when in regression equation.

Table 5. Measurement Model Statistics for Variables Used in the Path Mod

	Measurement Model Statistics						
Model	Est. Method	Chi-Square	df	RMSEA	CFI	GFI	AGFI
Emotional Stability	WLS	92.86	35	.09	.93	.97	.95
Extroversion	WLS	96.04	35	.09	.91	.97	.95
Organisational Citizenship Behaviour	WLS	127.82	35	.11	.94	.97	.95
Stress In General	WLS	232.31	89	.08	.85	.95	.93
Counter Productive Work Behaviour	WLS	106.29	35	.10	.86	.96	.94
Job Satisfaction	WLS	88.07	35	.08	.95	.97	.96

Note: N= 229; OCBs = organizational citizenship behaviors; WLS = weighted least squares; RMSEA = root mean square error of approximation; ECVI = expected cross validation index; CFI = comparative fit index; and GFI = goodness of fit index; AGFI = adjusted goodness of fit index

Discussion

In this analysis, we observed strong support for the criterion validity of the Big Five markers in New Zealand. Furthermore, the pattern of complex relations between two of Goldberg's IPIP scales (Extraversion and Emotional Stability) and psychological variables from the nomology of job satisfaction was found to be similar to overseas research, which provided evidence of the construct similarity between New Zealand and the USA, where the measure was originally developed.

Conclusions and Limitations

This research has found considerable support for the suitability of the Big Five personality markers in New Zealand. As markers for the FFM, this item set performs well, and in accordance with overseas research. This evidence came from multiple sources. First, support comes from examination of the dimensionality of the item set, which we found can be best represented by five factors. The five-dimensional structure showed the best fit in terms of many fit indices at both the item and parcel levels of analysis. If the dimensions claimed by Goldberg were not evident in the data, scale scores would be essentially meaningless, and, hence, our confidence in the extensive literature detailing Big Five dependent variable relations would be lessened in the New Zealand work context. The observed results, however, provide reason for confidence that the FFM structure is applicable in New Zealand. Secondly, results of measurement equivalence analyses for gender groups indicated that the questionnaire functions similarly across genders. The observed mean differences were also in accordance with a priori hypotheses based on overseas research findings. Finally, Goldberg's IPIP scales showed good predictive validity and two scales demonstrated hypothesized patterns of relations within the path model of antecedents and consequences of job satisfaction. The IPIP scales functioned in New Zealand as they did in Crede et al.'s (2003) study. Together, these results increase our confidence that other relations between Big Five variables and organizational outcomes observed overseas would also be observed in New Zealand. For applied use, however, we recommend the longer versions of these scales available at http://ipip.ori.org/

As with all applied research, there are weaknesses that must be identified and acknowledged in the hope they can be addressed by future researchers. The first such weakness is that these data were collected as part of a broader psychometric validation exercise, and, therefore, the participants were not job applicants. Research is equivocal with regard to the effect that the potential for faking has on the factor structure of personality questionnaire responses (e.g. Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). A second weakness is that the data were collected using standard self-report approaches. As a result, the size of the relationships is likely to be inflated by common method variance (Feldman & Lynch, 1988). A further limitation is that these analyses are cross-sectional, and as such, it is impossible to prove the causal direction of the observed relationships.

Future research should try to address these methodological limitations by examining the dimensionality of the FFM in New Zealand for job applicants, examining the relationship of the five factors to criteria assessed with alternative measures, and conducting longitudinal analyses to confirm our cross sectional inferences. Longitudinal designs facilitate detection of causality. There is also a pressing need to extend the measurement equivalence research to ethnic groups, and to extend the measurement equivalence to relational equivalence analyses. Only by examining relational equivalence, as well as measurement equivalence, can we be sure that not only are our tests and questionnaires measuring the same for interest groups, but that they are also predicting equivalently for all groups. While the cumulative research to date in the United States has not found evidence of differential prediction (Camilli & Shepard, 2000), this fact should not be taken for granted as being the case in New Zealand.

References

Anderson, J.C., & Gerbing, D.W. 1988. Structural Equation Modeling in Practice: A review and recommended two-step approach. *Psychological Bulletin, 103,*411-423.

Barrick, M. and Mount, M. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-25

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). The FFM personality dimensions and Job performance: Meta-Analysis of Meta-Analyses. Invited submission to a special "selection" issue of International *Journal of Selection and Assessment, 9,* 9-30.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Black, J. (2000). Personality testing and police selection: Utility of the 'Big Five'. *New Zealand Journal of Psychology, 29(1),* 2-9.

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences, 11,* 515-524.

Borman, W. C. & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10,* 99-109.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105,* 467-477.

Byrne (1998) *Structural Equation Modeling With Lisrel, Prelis, and Simplis: Basic Concepts, Applications, and Programming* (Multivariate Applications Book Series) Lawrence Erlbaum Associates.

Camilli, G. & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items.* Thousand Oaks, CA: Sage Publications.

Caprara, G. V., Caprara, M., & Steca, P. (2003). Personality's correlates of adult development and aging. *European Psychologist, 8,* 131-147.

Catell, R. B. (1947). Confirmation and clarification of primary personality factors. *Psychometrika, 12,* 197-220.

Chernyshenko, O.S., Stark, S., Credé, M., Wadlington, P., & Lee, W. (2003, April). *Improving the measurement of job attitudes: The development of the IJSI.* Paper presented at the 18th annual conference for the Society of Industrial and Organizational Psychologists. Orlando, Fl.

Costa, P. T., Teracciano, A., & McCrae, R. R. (2001). Gender Differences in Personality Across Cultures: Robust and Surprising Findings. *Journal of Personality and Social Psychology, 81,* 322-331.

Credé, M., Chernyshenko, O.S., Stark, S., Bashshur, M. R., Dalal, R. S., & Ben-Roy, D. (2003, April). *Development of an integrative model of the antecedents and consequences of job satisfaction.* Paper presented at the 18th annual conference for the Society of Industrial and Organizational Psychologists. Orlando, Fl.

Day, D. V., & Silverman, S. B. (1989). Personality and Job Performance: Evidence of Incremental Validity. *Personnel Psychology, 42,* 25-36.

Digman, J. M. (1997). Higher-Order Factors of the Big Five. *Journal of Personality and Social Psychology, 73,* 6, 1246-1256.

Drasgow, F. (1987). Study of the Measurement Bias of Two Standardized Psychological Tests. *Journal of Applied Psychology, 72,* 1, 19-29.

Eysenck, H. J. (1991). Dimensions of Personality: 16, 5, or 3? – Criteria for a taxonomic paradigm. *Personality and Individual Differences, 12,* 773-790.

Eysenck, H. J. (1992). Four ways five factors are *not* basic. *Personality and Individual Differences, 13,* 667-673.

Feldman, J., & Lynch, I. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology, 73,* 421-435.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116,* 429-456.

Gelatley, I. R., Paunonen, S. V., Meyer, J. P., Jackson, D. N., & Goffin, R. D. (1991). Personality, vocational interest, and cognitive predictors of managerial job performance and satisfaction. *Personality and Individual Differences, 12,* 221-231.

Goldberg, L. R. (1993). The Structure of Phenotypic Personality Traits. *American Psychologist, 48,* 1, 26-34.

Goldberg, L. R. (1990). An alternative "Description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59,* 1216-1229.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe, Vol. 7.* (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R. (in press). The comparative validity of adult personality inventories: Applications of a consumer-testing framework. In S. R. Briggs, J. M. Cheek, & E. M. Donahue (Eds.), *Handbook of Adult Personality Inventories.* New York: Plenum.

Hanisch, K. A., & Hulin, C. L. 1991. General attitudes and organizational withdrawal: An evaluation of a causal model. *Journal of Vocational Behavior, 39*, 110-128.

Harrison, D.A., McLaughlin, M.E., & Coalter, T.M. (1996). Context, cognition, and common method variance: psychometric and verbal protocol evidence. *Organizational Behavior and Human Decision Processes, 68,* 246-265.

Harter, J.K., Schmidt, F.L., & Hayes, T.L. (2002). Business unit-level relationship between employee satisfaction, employee engagement, and business outcomes: a meta analysis. *Journal of Applied Psychology, 87,* 2, 268-279.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, Fl: Academic Press.

Hough, L. & Ones, D. (2001). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In Anderson, D.S. Ones, K. Sinangil, & C. Viswesvaran (Eds.), *International handbook of work and organizational psychology.* Sage Publications.

Hough, L. M. (1998). Personality at work: Issues and evidence. In M. D. Hakel (Ed.), *Beyond Multiple Choice: Evaluating Alternatives to Traditional Testing for Selection* (pp.131-166). Mahwah, NJ: Lawrence Erlbaum Associates.

Hulin, C.L., & Judge, T.A. 2003. Job attitudes. In Borman, W.C., Ilgen, D.R., & Klimoski, R.J. (Eds.) *Handbook of Psychology, Volume 12* (pp.255-276). Hoboken: John Wiley & Sons.

Jöreskog, K. & Sörbom, D. (1986). *LISREL 6: Analysis of linear structural relationships by maximum likelihood and least square methods.* Mooresville, IN: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago, IL: Scientific Software.

Judge, T. A., Bono, J. E., & Locke, E. A. 2000. Personality and job satisfaction: The mediating role of job characteristics. *Journal of Applied Psychology, 85,* 237-249.

McCrae, R. R., Zonderman, M. H., Bond, M. H., Costa, P. T., Jr., & Paunonen, S. V. (1996). Evaluating Replicability of Factors in the Revised NEO Personality Inventory: Confirmatory Factor Analysis versus Procrustes Rotation. *Journal of Personality and Social Psychology, 70,* 552-566.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 57,* 17-40.

Motowidlo, S. J., Van Scotter, J. R., (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology, 79,* 4, 475-480.

Ones, D. S., & Viswesvaran, C. (1998) Gender, Age, and Race Differences on Overt Integrity Tests: Results Across Four Large-Scale Job Applicant Data Sets. *Journal of Applied Psychology, 83,* 1, 35-42.

Raju, N. S., Lafitte, L. J., & Byrne, B. M. (2002) Measurement Equivalence: A Comparison of Methods Based on Confirmatory Factor Analysis and Item Response Theory. *Journal of Applied Psychology, 87,* 3, 517-529.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552 – 566.

Saucier, G., & Goldberg, L. R. (2003). The structure of personality attributes. In M. R. Barrick & A. M. Ryan (Eds.), *Personality and work: Reconsidering the role of personality in organizations* (pp. 1-29). San Francisco, CA: Jossey-Bass.

Schmidt, F. L., & Hunter J. E. (1998). The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings. *Psychological Bulletin, 124,* 2, 262-274.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229 – 239.

Stanton, J. M., Balzer, W. K., Smith, P. C., Parra, L. F., & Ironson, G. (2001). A General Measure of Work Stress: The Stress in General Scale. *Educational and Psychological Measurement, 61,* 5, 866-887.

Stark, S., Chernyshenko. O., Chan, K., Lee, W., & Drasgow, F. (2001). Effects of the Testing Situation on Item Responding: Cause for Concern. *Journal of Applied Psychology, 86,* 5, 943-953.

Stark, S., Chernyshenko, O., & Drasgow, F. (2004). Examining the effects of differential item functioning and differential test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89,* 3, 497-508.

Stark, S., Chernyshenko, O., & Drasgow, F. (In review). Are CFA and IRT Equally Viable Methods for Detecting Biased Items? Toward a Unified Strategy for DIF Detection. *Journal of Applied Psychology.*

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioural Research, 25,* 173-180.

Tanaka, J. S., & Huba, G. J. (1984). Structures of Psychological Distress: Testing confirmatory hierarchical models. *Journal of Consulting and Clinical Psychology, 52,* 719-721.

Acknowledgements

This research was partially funded by Selector Ltd, and a grant from the New Zealand Department of Labour.

Address for correspondence:

Oleksandr Chernyshenko
Department of Psychology
University of Canterbury
Private Bag 4800
Christchurch, NZ.
Email:
sasha.chernyshenko@canterbury.ac.nz