Analysis of the Gender Variable in the Eysenck Personality Questionnaire Revised Scales Using Differential Item Functioning Techniques

Sergio Escorial and Maria J. Navas

Educational and Psychological Measurement 2007; 67; 990 originally published online Jun 6, 2007;
DOI: 10.1177/0013164406299108

Educational and Psychological Measurement
Volume 67 Number 6
December 2007 990-1001
© 2007 Sage Publications
10.1177/0013164406299108
http://epm.sagepub.com
hosted at
http://online.sagepub.com

Analysis of the Gender Variable in the Eysenck Personality Questionnaire–Revised Scales Using Differential Item Functioning Techniques

Sergio Escorial
Centro de Estudios Superiores Cardenal Cisneros
Maria J. Navas
Universidad Nacional de Educación a Distancia

Studies in the field of personality have systematically found gender differences in two of the three dimensions of the Eysenck model: neuroticism and psychoticism. This study aims to analyze these differences in the Eysenck Personality Questionnaire–Revised (EPQ-R) scales using differential item functioning (DIF) techniques to determine whether these differences are the result of a differential functioning of the items between males and females or if, on the contrary, they may be reflecting true differences in the assessed dimensions. To this end, 794 participants within a wide age range were evaluated using the EPQ-R test. The following detection methods were used in order to examine DIF: standardization, the simultaneous item bias test, logistic regression, Lord's χ^2 test, and the differential functioning of items and tests framework. According to the results, the gender differences observed do not seem to be the result of any flaw of the measuring instrument used.

Keywords: *EPQ-R; gender differences; DIF; standardization; logistic regression; Lord's χ^2 test; DFIT framework; SIBTEST*

The literature is rich in studies investigating gender differences in personality traits (e.g., see Feingold, 1994; Jorm, 1987). Gender differences are typically found in two traits: neuroticism and psychoticism. In the first case, it is women who score higher, normally between 0.25 and 0.50 standard deviations above the average for men. In the case of psychoticism, it is men who score higher, normally about 0.20 standard deviations. These differences appear to be a universal phenomenon because they are replicated regardless of the culture and period in which they are measured and regardless of the instrument used to measure these traits (Colom & Jayme-Zaro, 2004).

Authors' Note: Please address correspondence to Sergio Escorial, Centro de Estudios Superiores Cardenal Cisneros, División de Psicología, 28006 Madrid, Spain; e-mail: sergio.escorial@uam.es.

Gender differences have been found among adults, children, and young people in countries on five continents (Barrett & Eysenck, 1984; Delgado, 1995; Francis, 1993) using, among others, such distinct measuring instruments as the Eysenck Personality Inventory (Eysenck & Eysenck, 1975); the Eysenck Personality Questionnaire–Revised (EPQ-R; Eysenck & Eysenck, 1997); the Neuroticism, Extraversion, and Openness Personality Inventory–Revised (NEO PI-R; Costa & McCrae, 1992); and the Neuroticism Scale Questionnaire (Cattell & Sheier, 1961).

IQ is usually considered to be the best predictor of an individual's future performance in numerous socially relevant contexts (Jensen, 1998), but the use of standardized methods for the assessment of personality in selection processes is gradually increasing. In fact, high scores in neuroticism are sometimes being used in practice to reject individuals instead of to select them (Delgado, 1995). Similarly, high scores in psychoticism are not favorably regarded and sometimes become a criteria to exclude candidates from a selection process. Because tests are frequently used to make important decisions in people's lives, it is essential to determine if the differences observed between men and women in test scores are the result of real differences in personality; if they are the reflection of a diverse reality; or if, on the contrary, these differences are mere artifacts of the measuring instruments themselves, differences that can lead, for instance, to discrimination when men and women attempt to enter the labor and education markets. Differential item functioning (DIF) analysis can help to examine this question.

An item is said to display DIF when the probability that participants matched at the measured underlying trait will choose a particular response alternative depends on the group they belong to; the probability is not the same for participants in different groups with the same level for the measured trait. An item is said to display impact when the probability of choosing a particular response differs from one group to another, that is, when there are differences in the average performance of the groups for the item.

Most studies conducted in the field of personality have limited themselves to the assessment of impact. Studies on the existence of possible DIF in the tests commonly used to assess personality are scarce and recent (Borsboom, Mellenbergh, & Van Heerden, 2002; Collins, Raju, & Edwards, 2000; Ellis & Mead, 2000; Gelin & Zumbo, 2003; Lange, Irwin, & Houran, 2000; Reise, Smith, & Furr, 2001; Smith, 2002). The main goal of this investigation was to analyze gender differences using DIF techniques in the traits noted in the Eysenck model. To this end, we began by assessing the impact and then the possible existence of DIF in the EPQ-R test by means of a battery of DIF detection techniques.

Method

Instruments

The Spanish version of the EPQ-R was used, a test consisting of 83 dichotomous items assessing the following traits: Extraversion, Neuroticism, and Psychoticism,

Table 1
Reliability Estimates of the Eysenck Personality Questionnaire–Revised Test

	Extraversion		Neuroticism		Psychoticism	
	Males	Females	Males	Females	Males	Females
α	.86	.83	.87	.86	.76	.73
95% CI	.84–.88	.80–.85	.85–.89	.83–.88	.73–.80	.70–.76

Note: CI = confidence interval.

plus an additional Dissimulation scale. Table 1 shows the values obtained in the present study for the Cronbach's coefficient α based on gender for the Extraversion, Neuroticism, and Psychoticism scales, and the 95% confidence interval for these reliability estimates using a central F distribution approach (Fan & Thompson, 2001).

These estimates were very similar for the male and female groups and exceeded the generally accepted .80 cutoff value for general research purposes (Henson, 2001; Nunnally & Bernstein, 1994), as far as the Extraversion and Neuroticism scales were concerned. The lower reliability coefficients obtained in the Psychoticism scale were consistent with the results yielded by other studies (Eysenck & Eysenck, 1997; Ortet, Ibáñez, Moro, Silva, & Boyle, 1999).

Participants

The sample consisted of 794 participants selected from a target population aged 18 and older, by means of a gender and age quota sampling technique. Data were collected by 75 evaluators (mostly members of the research team and graduate students) instructed to administer the test to one man and one woman in each of the following age subgroups: < 20, 20-29, 30-39, 40-49, 50-59, and > 60. Fieldwork was carried out for 3 weeks, and the nonresponse rate (approached participants who refused to participate) was 12%. The final gender distribution of the sample was composed of 55% male participants and 45% females, with a similar gender percentage in each of the six age subgroups.

Analysis

The relationship between gender and item responses was examined in order to detect impact. For the global scales, impact was evaluated by means of a statistical test of the difference between means of two independent samples. Furthermore, an effect size measure (Cohen's d) was calculated both at the item and at the scale level.

The techniques used to detect DIF were logistic regression, standardization, Lord's χ^2 test, differential functioning of items and tests (DFIT) model-based

statistics and the simultaneous item bias test (SIBTEST). Given the fact that the Lord's χ^2 test and the DFIT model-based statistics operate within the framework of item response theory (IRT), it was necessary to verify that the data obtained showed a reasonable fit to an IRT model before either of these statistics was calculated. To this end, the approach suggested by Hambleton and Swaminathan (1985) was followed: The assumptions underlying the models under consideration were checked, and the goodness of fit and the invariance of the parameters of the model were assessed. The IRT model that showed the best fit to the data was the two-parameter logistic model, although it was necessary to remove an item in the Neuroticism scale (Item 72) and seven items in the Psychoticism scale (Items 9, 50, 51, 63, 66, 71, and 80). Item 55 in this latter scale was also removed because the number of participants responding in the direction of the trait was clearly insufficient to carry out DIF analysis (17 out of 794).

Logistic regression. When logistic regression is used to study DIF (Swaminathan & Rogers, 1990), the goal is to determine if it is sufficient to introduce the level of ability of the participants in the mathematical function to predict the item responses (model without DIF) or if, on the contrary, it is necessary to include a term that refers to the group membership of the participant under consideration (uniform DIF model) or a term that refers to the interaction between the participant's group membership and his or her ability (nonuniform DIF model). We speak of uniform DIF whenever the probability of choosing a particular response option is systematically greater for one group of examinees throughout the trait continuum; otherwise, it would be nonuniform DIF.

Zumbo and Thomas (1997) proposed combining the statistical significance with a measure of the effect size in order to conclude whether an item displayed DIF, using a model comparison strategy. The statistical significance was defined by the χ^2 difference of the compared models. As far as the practical significance was concerned, an effect was considered to exist when the increase in the squared multiple correlation coefficient in the compared models was at least .035. Thus, values between .035 and .070 would indicate a moderate DIF and values above .070 a large DIF (Jodoin & Gierl, 2001). This was the strategy used in this study.

Standardization. The numerical index used to quantify uniform DIF is the standardized difference in proportions (STD P-DIF; Dorans & Holland, 1993):

$$STD\ P-DIF = \sum_{j=1}^{J} \frac{W_j}{\sum_{j=1}^{J} W_j}(P_{F_j} - P_{R_j}), \quad (1)$$

where W_j is the weighting factor at the score level j, normally the number of individuals in the minority or the socially disadvantaged group (commonly referred as the focal group) at that score level; and P_{F_j} and P_{R_j} are the proportions of individuals

who respond to the item in the direction of the trait at the j score level in the focal group and in the reference group, respectively.

The software used was the Dimensionality-Based DIF/DBF Package (Stout & Roussos, 1999).

Lord's χ^2 statistic. This statistic is based on IRT and tests the null hypothesis that the parameters defining the item characteristic curve are the same for the focal and for the reference groups or, in other words, that the item does not display DIF (Lord, 1980). The statistic's value was determined using the LINKDIF program (Waller, 1998). This program also performed the previous equating of the estimates obtained for the item parameters in both groups. These estimates were obtained using the MULTILOG program (Thissen, 1991).

The DFIT model. The DFIT model is the most recent approach used in this study. It was proposed by Raju, van der Linden, and Fleer (1995) within the framework of IRT. As its own name suggests, the DFIT model makes it possible to study not only DIF but also the differential functioning of the overall test. In particular, this model contemplates a differential test functioning statistic (DTF) and two DIF statistics, a compensatory index (CDIF) and a noncompensatory index (NCDIF):

$$CDIF_i = Cov(d_i, D) + \mu_{d_i}\mu_D \tag{2}$$

$$NCDIF_i = \sigma^2_{d_i} + \mu^2_{d_i} \tag{3}$$

$$DTF = \sigma^2_D + \mu^2_D = \sum_{i=1}^{n} CDIF_i, \tag{4}$$

where d_i is the difference in the probability of responding in the direction of the trait to item i in the focal and reference groups, and D is the difference between the test's expected true scores for both groups.

There exist statistical significance tests for the DTF and NCDIF indices but not for the CDIF index. In this case, we proceeded in the following manner to interpret the values obtained. If the global index was not statistically significant, it was assumed that none of the CDIF values was statistically significant; however, if the DTF index was statistically significant, then the items displaying a higher absolute CDIF value were removed one by one until the DTF index ceased to be statistically significant. When this occured, the items removed were considered to display DIF.

It is also essential to introduce a practical significance consideration in the interpretation of all these DFIT statistics, whether an associated statistical significance test is available or not. The reason is that, in the first case, the tests depend to a large extent on the available sample size, which causes some NCDIF statistic

values that are practically null to become statistically significant. In the second case, it has been further observed that CDIF values very close to zero also result in items labeled as displaying DIF. The following strategy was used in the present study: Mark as DIF items those that the previous iterative procedure identified as such (in the case of CDIF), or those revealed by the statistical significance test (in the case of NCDIF), but only if the value obtained for the statistic (CDIF or NCDIF) was higher than a cutoff value. Those values were obtained through the simulation of the sampling distribution of both statistics working with non-DIF conditions and with 4,000 values, and the cutoff values corresponded to the 99.5th percentile of the distribution. A modified version of the LINKDIF program was used to obtain all these indices and to perform the corresponding significance tests.

SIBTEST. This procedure detects differential item and test functioning (Shealy & Stout, 1993). In particular, the averages for the focal and for the reference groups are compared in a subtest consisting of a set of items suspected to be likely to exhibit DIF (suspect subtest), matching the participants according to the score obtained in another subtest consisting of test items considered not to present DIF (valid subtest). The following statistic, along with its corresponding test of statistical significance, is calculated as

$$\beta = \sum_{j=0}^{m} W_j \left(\overline{Y}_{R_j} - \overline{Y}_{F_j} \right), \tag{5}$$

where W_j is the weighting factor used at the ability level j, usually defined as the proportion of participants in the focal group with a j score in the valid subtest; and $\overline{Y}_{R_j}$ and $\overline{Y}_{F_j}$ are the averages in the suspect subtest of the participants with a score equal to j in the valid subtest, for the reference and the focal groups, respectively.

In the present study, the information obtained using the DIF detection procedures previously described was used to decide which items were in the valid subtest and which were in the suspect subtest in each scale. The analyses were performed using the Dimensionality-Based DIF/DBF Package program (Stout & Roussos, 1999).

Results

Impact

Table 2 summarizes the results obtained when impact at the item level was evaluated, indicating the items that displayed a statistically significant relationship ($p < .01$) between the gender and the item response. The table also shows which group scored higher and the effect size range (d) of those items displaying a statistically significant impact.

Table 2
Items With Impact

Scale	In Favor of Males	In Favor of Females	*d* Range
Extraversion	28, 49	46, 57	.174–.265
Neuroticism	—	2, 4, 13, 19, 24, 32, 35, 41, 42, 52, 54, 62, 75, 76, 78, 81	.201–.450
Psychoticism	15	—	.304

These results indicate that the Neuroticism scale presented a very high proportion of items with impact in favor of women. Regarding the Extraversion scale, few items with impact were detected, and these were equally distributed between the male and the female groups. Finally, only one item with impact in favor of men was detected in the Psychoticism scale. The effect size of the differences found in the three scales was rather small. The only scale in which statistically significant differences in the total scores were found was the Neuroticism scale ($t = -6.974$, $p < .001$, and $d = .498$).

DIF

To establish whether these differences persisted or disappeared when comparing males and females previously matched at their level of neuroticism (extraversion or psychoticism), the relevant analyses were performed to detect the possible differential functioning. Table 3 shows the results obtained when the previously described methods were applied to the items of the three EPQ-R scales, with the exception of SIBTEST. The items included in this table are those the analyses revealed to be statistically significant ($p < .01$), indicating when appropriate if there was DIF in favor of the male or female group.

When using the logistic regression, DIF was detected in four items on the Extraversion scale, but only Item 46 approached the established moderate DIF level; this was also the only item with nonuniform DIF. In the other scales, only one item displayed uniform DIF, but the effect size was very small for the item on the Neuroticism scale, barely approaching the level considered moderate on the Psychoticism scale. The results obtained with the standardization method showed that there were four items with uniform DIF on the Extraversion scale, three on the Neuroticism scale, and one on the Psychoticism scale.

To determine the type of DIF displayed when working with IRT-based methods, the exact area measures defined by Raju (1988, 1990) were examined with the LINK-DIF program, and the corresponding characteristic curves of the DIF items were plotted for the focal and the reference groups. Three DIF items were detected on the Extraversion scale with the NCDIF index and seven items with the χ^2 statistic. As

Table 3
Items With Possible Differential Item Functioning (DIF) Detected by Different Procedures

	Logistic Regression			Standardization		χ^2			DFIT Model					
									CDIF			NCDIF		
	DIF U			DIF U		DIF U			DIF U			DIF U		
	M	F	DIF NU	M	F	M	F	DIF NU	M	F	DIF NU	M	F	DIF NU
Extroversion	28, 49	57	46	49	3, 46, 57	28, 49, 70	3, 46, 57	69	—	—	—	49	46	69
Neuroticism	8	—	—	8, 18	81	8, 18	13, 81	83	—	—	—	8, 18, 64	—	—
Psychoticism	15	—	—	15	—	15	44	—	29	—	—	—	—	—

Note: DFIT = differential functioning of items and tests; CDIF = compensatory DIF index; NCDIF = noncompensatory DIF index; DIF U = uniform DIF model; DIF NU = nonuniform DIF model; M = male; F = female.

shown in Table 3, the type of DIF was uniform in practically all the items. Only Item 69 exhibited nonuniform DIF. When operating with the NCDIF index on the Neuroticism scale, there were three items with DIF. The number went up to five when the χ^2 statistic was applied; it was basically uniform DIF, with the exception of Item 83. On the Psychoticism scale, some items with uniform DIF were also detected, two with the χ^2 statistic and one with the CDIF index. When the DTF was evaluated in each of the scales, it was observed that none of the three functioned differentially.

Finally, Table 4 succinctly presents the most relevant information concerning the analyses carried out using the SIBTEST procedure. This table shows the number of items in the valid subtest and the items of which the suspect subtest for each of the scales consisted. The suspect subtest consisted of those items identified as DIF items by logistic regression and standardization, or by one of these procedures and an IRT-based index. The reason why the consistency of the results obtained using non IRT-based methods was emphasized was related to the fact that the study sample size was not excessively large, even though it was large enough to operate within IRT. The table also shows which items displayed uniform DIF and which displayed nonuniform DIF. Last, on each scale it is indicated if the differential suspect subtest functioning was or was not statistically significant.

Three items displayed DIF on the Extraversion and Neuroticism scales; two of them displayed uniform DIF and the third nonuniform DIF. No differential functioning was detected in the suspect subtest on either scale. This was due to the cancellation effect, as there were items on both scales that favored in one case the female group and in another case the male group. On the contrary, on the Psychoticism scale Item 15 was flagged for DIF and the scale showed differential functioning. Nevertheless, given the fact that the number of participants who responded affirmatively to this item was rather low (110 males and 48 females out of 794 participants), it is necessary to interpret these data cautiously and to try to replicate them in a much larger sample where the number of males and females responding in the two possible directions will allow for an appropriate application of the DIF detection techniques.

If Tables 3 and 4 are examined jointly, it becomes evident that there was a remarkable consistency in the results obtained when using different procedures to detect DIF. Thus, out of the nine items of which the suspect subtests of the three scales consisted, six items were flagged for DIF by at least four out of the five procedures used (Items 8, 15, 18, 46, 49, and 57), and one was flagged by means of three procedures (Item 81). In only two items on the Extraversion scale (Items 3 and 28) was the presence of DIF detected by just two procedures.

Discussion

The results obtained show that the only scale in which significant differences were found was the Neuroticism scale, both in the global scores and in a fairly

Table 4
Results Obtained Using the Simultaneous Item Bias Test Procedure

Scale	*N* of Items in the Valid Subtest	Items in the Suspect Subtest	Uniform DIF in Favor of		Nonuniform DIF	DTF (*p*)
			Males	Females		
Extraversion	14	3, 28, 46, 49, 57	49	57	46	.183
Neuroticism	19	8, 18, 81	18	81	8	.147
Psychoticism	14	15	—	—	15	.001

Note: DIF = differential item functioning; DTF = differential test functioning.

large number of the items in the scale. The direction of the differences was consistent with the existing literature: Women scored higher in neuroticism.

Nevertheless, this study did not confirm the differences typically found in psychoticism. This lack of consistency with the existing literature could be related to the age variable: The differences in psychoticism between men and women may not be constant throughout life. Thus, whereas in adolescence and in early adulthood it is men who score higher, in later stages these differences may dissipate, as the scores in psychoticism decrease with age for men but remain stable for women (Colom & Jayme-Zaro, 2004). Most studies work with samples within a very homogeneous age group close to early adulthood. This study, however, analyzed a very wide age range.

The subsequent study of the possible differential functioning showed that the differences initially revealed by the analysis of the impact were not a consequence of problems derived from the measuring instrument. In effect, DIF was detected in the EPQ-R test in only a few items, for which the effect size was, moreover, moderate at best. In addition, a cancellation effect could be further observed, implying that the scale overall did not function differently for the male and the female groups.

In view of these results, the interesting question arose of whether the items that functioned differentially in the present study were at least partially responsible for the differences found between males and females when impact was assessed. To answer this question, the items of the valid subtest used in the SIBTEST procedure were selected. That is, the items selected were those that can be assumed with justifiable certainty not to function differentially depending on gender and that provided a valid measure of the trait assessed according to the inclusion criteria previously discussed. Next, the average score for males and females for each of the scales was calculated in the valid subtest, and the means were compared. The results showed that the tendencies did not change when items displaying DIF were removed. In other words, in those scales where significant differences between males and females existed, these persisted; whereas in the scales where such differences did not exist, they continued not to exist. Furthermore, the effect sizes were similar in the original and in the reduced scale.

In sum, the present study brings empirical evidence to support the fact that differences in the personality test scores on the EPQ-R are not the result of biases in the measuring instrument. These results have important implications for the validity of scores on this test, as they represent one more step in the process of gathering favorable evidence in support of the use of scores obtained in applied contexts.

References

Barrett, P., & Eysenck, S. (1984). The assessment of personality factors across 25 countries. *Personality and Individual Differences*, *5*, 615-632.

Borsboom, D., Mellenbergh, G., & Van Heerden, J. (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement*, *26*, 433-450.

Cattell, R., & Scheier, I. (1961). *Handbook for the Neuroticism Scale Questionnaire: The NSQ.* Champaign, IL: IPAT.

Collins, W., Raju, N., & Edwards, J. (2000). Assessing differential functioning in a Satisfaction scale. *Journal of Applied Psychology*, *85*, 451-461.

Colom, R., & Jayme-Zaro, M. (2004). *La psicología de las diferencias de sex* [Psychology of sex differences]. Madrid: Biblioteca Nueva.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory and NEO Five-Factor Inventory (NEO-FFI).* Obessa, FL: Psychological Assessment Resources.

Delgado, C. (1995). Sesgo de género en la medición del neuroticismo [Gender bias in neuroticism measurement]. *Ciencias Sociales*, *69*, 51-66.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Ellis, B., & Mead, A. (2000). Assessment of the measurement equivalence of a Spanish translation of the 16PF questionnaire. *Educational and Psychological Measurement*, *60*, 787-807.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire.* London: Hodder & Stoughton.

Eysenck, H. J., & Eysenck, S. B. G. (1997). *Cuestionario revisado de personalidad de Eysenck (EPQ-R)* [Manual of the Eysenck Personality Questionnaire-Revised].Madrid: TEA Ediciones.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement*, *61*, 517-531.

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

Francis, L. (1993). The dual nature of the Eysenckian Neuroticism scales: A question of sex differences? *Personality and Individual Differences*, *15*, 43-59.

Gelin, M., & Zumbo, B. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the Center for Epidemiologic Studies Depression Scale. *Educational and Psychological Measurement*, *63*, 65-74.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189.

Jensen, A. (1998). *The g factor.* London: Praeger.

Jodoin, M., & Gierl, M. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*, 329-349.

Jorm, A. (1987). Sex differences in neuroticism: A quantitative synthesis of published research. *Australian and New Zealand Journal of Psychiatry*, *21*, 501-506.

Lange, R., Irwin, H., & Houran, J. (2000). Top-down purification of Tobacyk's revised Paranormal Belief scale. *Personality and Individual Differences*, *29*, 131-156.

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Ortet, G., Ibáñez, M., Moro, M., Silva, F., & Boyle, G. (1999). Psychometric appraisal of Eysenck's Psychoticism scale: A cross cultural study. *Personality and Individual Differences*, *27*, 1209-1219.

Raju, N. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495-502.

Raju, N. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*, 197-207.

Raju, N., van der Linden, W., & Fleer, P. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*, 353-368.

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

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-240). Hillsdale, NJ: Lawrence Erlbaum.

Smith, L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin*, *28*, 754-763.

Stout, W., & Roussos, L. (1999). Dimensionality-based DIF/DBF package [Computer software]. Urbana-Champaign: William Stout Institute for Measurement, University of Illinois.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.

Thissen, D. (1991). *MULTILOG: Multiple, categorical item analysis and tests scoring using item response theory*. Chicago: Scientific Software.

Waller, N. (1998). LINKDIF: Linking item parameters and calculating IRT measures of differential item functioning of items and tests. *Applied Psychological Measurement*, *22*, 392.

Zumbo, B., & Thomas, D. (1997). *A measure of effect size for a model-based approach for studying DIF*. Prince George, BC, Canada: Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia.