



LUND UNIVERSITY

Using the Johnson-Neyman Procedure to Detect Item Bias in Personality Tests A Proposed New Method and Practical Guidelines for Data Analysis

Tunca, Burak

Published in:
The Wiley Handbook of Personality Assessment

DOI:
[10.1002/9781119173489.ch25](https://doi.org/10.1002/9781119173489.ch25)

2016

Document Version:
Peer reviewed version (aka post-print)

[Link to publication](#)

Citation for published version (APA):
Tunca, B. (2016). Using the Johnson-Neyman Procedure to Detect Item Bias in Personality Tests: A Proposed New Method and Practical Guidelines for Data Analysis. In U. Kumar (Ed.), *The Wiley Handbook of Personality Assessment* (pp. 346-360). Wiley-Blackwell. <https://doi.org/10.1002/9781119173489.ch25>

Total number of authors:
1

General rights

Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: <https://creativecommons.org/licenses/>

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

**Using the Johnson-Neyman Procedure to Detect Item Bias in Personality Tests:
A Proposed New Method and Practical Guidelines for Data Analysis**

Burak Tunca

University of Agder, Norway

Author Note

This is the author's version of the following book chapter: Tunca, B. (2016) Using the Johnson-Neyman Procedure to Detect Item Bias in Personality Tests, in The Wiley Handbook of Personality Assessment (ed U. Kumar), John Wiley & Sons, Ltd, Chichester, UK., which has been published in final form at <http://onlinelibrary.wiley.com/doi/10.1002/9781119173489.ch25/summary>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

**Using the Johnson-Neyman Procedure to Detect Item Bias in Personality Tests:
A Proposed New Method and Practical Guidelines for Data Analysis**

Personality researchers are often interested in examining trait differences between groups. For instance, are men more assertive than women? Or, are the Americans more impulsive than the Chinese? The common practice in answering such questions is first to administer the same personality scale to members of each group, and then to compare groups' scores on the scale. Validity of such comparisons, however, rests on the assumption that the scale items are not biased: Respondents with different group memberships understand and interpret the scale items in a similar manner. If this assumption is violated, the validity of the results becomes questionable. Personality researchers have long been warned against this potential problem in between-group comparisons (e.g., Smith, 2002; Thissen, Steinberg, & Gerrard, 1986).

The significance of this problem has also led to development of various statistical techniques to detect biased items (for reviews, see Reynolds, 2000; Zumbo, 2007). In the current chapter, I first review one of these techniques known as the analysis of variance (ANOVA) procedure (van de Vijver and Leung, 1997), which has been widely used in personality research (e.g., Caprara, Barbaranelli, Bermúdez, Maslach, & Ruch, 2000; Ramírez-Esparza, Gosling, Benet-Martínez, Potter, & Pennebaker, 2006; Vecchione, Alessandri, & Barbaranelli, 2012). Next, I propose an alternative to the ANOVA method. The alternative method, which is based on the Johnson-Neyman procedure (Johnson & Neyman, 1936), has the potential to overcome some of the major weaknesses of the ANOVA procedure. I introduce the proposed method in a non-technical manner and I present practical guidelines for data analysis using an add-on for mainstream statistical software packages (PROCESS Macro; Hayes, 2013), so that researchers

who are inexperienced in item bias analysis can easily apply the Johnson-Neyman procedure to their research.

What is Item Bias?

Item bias, also known as *differential item functioning (DIF)*, refers to item level anomalies within an instrument that can threaten the validity of group comparisons. In psychology research, an item is considered to be biased when respondents with different group memberships score differently on a survey item, while being at the same level of the latent trait (Smith, 2002; van de Vijver & Leung, 2011). To illustrate, Santor, Ramsay, and Zuroff (1994) examined gender level item bias in the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and found that men and women, who were equally depressed, responded differently to an item related to their perceived body image distortion: At all levels of depression, women were more likely than men to report concerns about looking unattractive. The authors concluded that the item was endorsed differently across gender groups, thus it was biased, and scores for this item would lead to misleading results when comparing the degree of depression between men and women.

Item bias exists in two forms: *Uniform* and *non-uniform* (van de Vijver & Leung, 1997). Uniform bias manifests itself when there is a systematic difference between groups for an item score across all score levels (see figure 26.1-a). This indicates that one group endorses the item differently than the other group (either at a higher or lower level), regardless of their total score on the latent trait. Non-uniform bias, on the other hand, is a less common form, and it occurs when the item score differences between groups are not systematic across all score levels (see figure 26.1-b). For example, an item may be more discriminating for one group at lower score

levels and more discriminating for the other group at higher score levels, which implies an interaction effect of group membership and total score on the latent trait. Various statistical techniques have been proposed to test for evidence of uniform and non-uniform item bias. While item response theory (IRT) and logistic regression methods are commonly used to detect item bias in dichotomous scores, the ANOVA procedure (van de Vijver and Leung, 1997) has been a popular technique in examining item bias in numerical scores.

<FIGURE 26.1 HERE>

The ANOVA Procedure

To test for evidence of item bias in unidimensional scales, van de Vijver and Leung (1997) introduced a procedure based on conditional analysis of variance (ANOVA). To illustrate how the ANOVA procedure functions, consider a test instrument with ten items (*item 1* to *item 10*) measured with a seven-point Likert-type scale (1 = strongly disagree to 7 = strongly agree), which was administered to members of two groups (*group A* and *group B*) to examine between-group differences. The ANOVA procedure is centered on three variables (i.e., *item*, *group*, and *score level*): *Group* and *score level* are categorical independent variables and *item* is a continuous dependent variable.

Item is the dependent variable in the ANOVA procedure, and it refers to the item we would like to examine for evidence of bias in the research instrument. In our hypothetical example there are ten items (*item 1* to *item 10*), and in the ANOVA procedure each item is examined independently for evidence of bias. *Group* and *score level* are the independent variables. *Group* refers to the groups that are compared in the study (e.g., gender or culture; in

our example *group A* and *group B*). The *group* variable is dummy coded in the dataset (e.g., *group A* = 0 and *group B* = 1).

Unlike *item* and *group*, which are already present in the dataset, *score level* is a new variable that must be calculated by the researcher. This is done by first computing a *total score* variable, which is simply the sum of all item scores (i.e., $item\ 1 + item\ 2 + \dots + item\ 10$) for each respondent. Recall that the test instrument had ten items measured by a seven-point scale. If a respondent selects 1 (strongly disagree) for all items, she will have the lowest possible total score (i.e., $10 \times 1 = 10$). On the other hand, if she selects 7 (strongly agree) for all items, her total score will be the maximum possible score (i.e., $10 \times 7 = 70$). Thus, each respondent's total score on the test instrument is a value between 10 and 70. Next, the *total score* variable is transformed into the *score level* variable. *Score level* is a categorical variable that is created by splitting the continuous *total score* variable into groups based on predetermined cut-points. The aim of this procedure is to group respondents with similar total scores, so that groups are ranging from "low total scores" to "high total scores". According to van de Vijver and Leung (1997), the cut-points should be determined in a way that each score level contains at least 50 respondents. If the sample size is 500, for instance, there are eight or nine groups in the *score level* variable (for a more detailed explanation of this procedure, see van de Vijver & Leung, 1997). It should be noted that these cut-points, which designate score levels, are arbitrary.

Once the three variables (i.e., *item*, *group*, and *score level*) are ready, a conditional two-way ANOVA is conducted to test for item bias. Conceptually, the ANOVA procedure tests the null hypothesis that "item scores are not different between groups" across different score levels. The ANOVA identifies three effects on the dependent variable *item*: The main effects of *score level* and *group*, and the *group x score level* interaction effect. The main effect of *score level* is

of little interest to the researcher; respondents at higher score levels usually have higher score on the item than respondents at lower score levels do, thus this main effect will often be significant. The significance levels of the remaining two effects are, however, essential to the ANOVA procedure.

When both the main effect of *group* and the *group x score level* interaction effect are non-significant, the item under scrutiny is unbiased (van de Vijver & Leung, 1997). On the other hand, a significant main effect of *group* and *group x score level* interaction effect indicate the presence of uniform and non-uniform bias, respectively (van de Vijver & Leung, 1997). Given our example, if respondents in *Group A* score higher (or lower) than respondents in *Group B* for *item 1* at all score levels, this uniform bias for *item 1* will be evident as a significant main effect of the *group* variable. Finally, if the differences between groups are not equivalent across all score levels, this non-uniform bias will be evident as a significant *group x score level* interaction effect.

The ANOVA procedure is easy to apply and interpret using mainstream statistical software packages. It also enables visual examination of item bias with graphical displays. Despite such advantages, the ANOVA procedure has two interrelated shortcomings: The need to discretize a continuous variable (i.e., transforming the continuous *total score* variable into the categorical *score level* variable) and the need for large sample sizes. Given that the higher the number of score levels, the more sensitive the analysis of item bias (van de Vijver & Leung, 1997, 2011), large sample sizes become necessary for the ANOVA procedure to achieve the “minimum 50 respondents per score level” rule of thumb and to retain the statistical power reduced as a result of the discretization procedure. Large sample sizes may, however, inflate significance levels, thereby making statistical significance testing less informative. Recognizing

this issue, Van de Vijver and Leung (2011) suggest using effect sizes instead of significance values to detect biased items with the ANOVA procedure.

There is compelling evidence in the literature concerning the limitations of discretizing a continuous variable based on arbitrary groups (e.g., Fitzsimons, 2008; Irwin & McClelland, 2003; MacCallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delaney, 1993). Taken together, these studies conclude that the discretization procedure should be avoided as it may result in reduced statistical power, loss of information, and misleading results. We can therefore conclude that the main shortcoming of the ANOVA procedure stems from discretizing a continuous variable to create a categorical alternative. Nonetheless, there are alternative methods such as the Johnson-Neyman procedure (Johnson & Neyman, 1936), which allows conducting similar analyses without discretizing continuous variables.

The Johnson-Neyman Procedure

The Johnson-Neyman procedure, introduced by Johnson and Neyman (1936) and extended to multiple regression models by Preacher, Curran, and Bauer (2006), has been proposed as an alternative to analysis of covariance techniques (D'Alonzo, 2004; Miyazaki & Maier, 2005). Analogous to the ANOVA method, The Johnson-Neyman procedure is used for examining the conditional effect of an independent variable on a dependent variable at different values of another independent variable (i.e., a moderator). Unlike the ANOVA method, the Johnson-Neyman procedure does not require any transformation of the continuous variables. Instead of discretizing a continuous variable into groups, it tests the same null hypothesis that “item scores are not different between groups” at all levels of the continuous variable and returns a “region of significance” within which the scores between groups differ at a specified

significance level (e.g., $p < .05$). The region of significance can then be plotted for a visual inspection of the results.

Based on its similarities with the ANOVA method, the Johnson-Neyman procedure can easily be adopted to item-bias analysis. In the item bias analysis context, the Johnson-Neyman procedure is also based on three main variables (i.e., *item*, *group*, and *total score*), which were explained in the previous section. Note that the categorical *score level* variable is redundant for the Johnson-Neyman procedure as the continuous *total score* variable can be used without any transformation. These three variables can be entered into a moderated multiple regression analysis as follows: *Item* is the dependent variable, whereas *group*, *total score*, and *group x total score* interaction are the independent variables. The multiple regression analysis then provides an output with two main effects (*group* and *total score*) and one interaction effect (*group x total score*). As for the ANOVA method, the significant main effect of *group* indicates uniform bias and the significant *group x total score* interaction effect indicates non-uniform bias. At this point, the region of significance can also assist the researcher in determining biased items such that high proportions of the sample within the significance region provide further evidence of item bias.

The biggest advantage of the Johnson-Neyman procedure over the ANOVA method in detecting item bias is that it accommodates the *total score* variable in the analysis as a continuous variable and tests the conditional effect of *group* on *item* at all levels of the *total score* variable. Testing all levels of the *total score* variable is more informative to the researcher, because respondents' total item scores are arbitrary and there are no meaningful values that can be selected and analyzed independently (see also floodlight analysis by Spiller, Fitzsimons, Lynch, & McClelland, 2013). For example, when conducting the Johnson-Neyman analysis with

the body mass index (BMI) as the moderator (e.g., Spiller et al., 2013), researchers may choose to examine the moderator at specific meaningful values (e.g., overweight respondents; BMI > 25). The *total score* variable in the item bias analysis, however, does not have such focal meaningful values; it therefore needs to be analyzed at all levels. Treating the *total score* variable as continuous rather than discretizing it with arbitrary cut-points also avoids loss of statistical power and possibility of spurious effects (West, Aiken, & Krull, 1996), which may be of concern when the ANOVA method is used.

Although the Johnson-Neyman procedure is not a new development, its complexity and lack of availability in statistical analysis programs has impeded its implementation. Today, however, add-ons like PROCESS Macro (Hayes, 2013) enable researchers to conduct the Johnson-Neyman analysis with ease using mainstream statistical packages (e.g., SPSS and SAS).

An Illustrative Example Using Conscientiousness

An illustration of the Johnson-Neyman procedure is presented here using a dataset that was collected from a university student sample in Norway. 230 students (120 females and 110 males) responded to a brief conscientiousness measure (8 items; Saucier, 1994) using a 7-point Likert-type scale (1 = strongly disagree to 7 = strongly agree) as part of a battery of survey questions. Our objective in this example is to examine item bias across gender groups. The scale had satisfactory alpha levels ($\alpha_{\text{total}} = .84$, $\alpha_{\text{male}} = .83$, $\alpha_{\text{female}} = .83$). Because the purpose here is to illustrate the procedure, advanced unidimensionality tests were not conducted and the scale was assumed to be unidimensional. Researchers are recommended to ensure their instruments' unidimensionality prior to conducting item bias analysis.

The results of an independent samples t-test analysis showed significant differences for various individual item scores and total scale mean score, such that females scored higher on trait conscientiousness than males did (see table 26.1). An examination of effect sizes (Cohen's *d*) in table 26.1 shows that the magnitude of gender differences across items vary from small to medium (Cohen, 1992). It is important to take note of this variation in the effect sizes, as it may signal presence of item bias in the scale (Smith, 2002).

<TABLE 26.1 HERE>

To conduct the main item-bias analysis, first the total score variable (i.e., *total*) was created. Recall that this variable is the sum of all item scores for each respondent. Hayes's PROCESS Macro for SPSS (Hayes, 2013) was used for data analysis (information regarding the installation of the macro is available online at <http://www.processmacro.org>). Separate analyses were conducted for each of the eight items in the scale. In the main PROCESS Macro dialogue box, first the scale item of interest (e.g., *careless*) was entered into the outcome variable (*Y*) box. Next, the dummy coded group variable (i.e., *gender*) was entered into the independent variable (*X*) box. Finally, the total score variable (i.e., *total*) was entered into the moderator variables (*M*) box. Model number 1 (the moderated regression model) was also selected in the main screen menu. In the "options" menu, "mean center for products" option was selected. This option mean centers the independent variables (*X* and *M*; *gender* and *total*) prior to computation of the interaction variable (*X x M*; *gender x total*). Mean centering renders the regression coefficients of the independent variables more meaningful (Hayes, 2013); given that our main interest lies in interpreting the main effect of the group variable (in our case *gender*), mean centering must be

employed. Graphical representations of the results are useful in item bias analysis; hence, the “generate data for plotting” option in the “options” menu was also selected. The final step was to select the “Johnson-Neyman” option from the menu titled “conditioning”. Pressing the “OK” button in the main dialogue box ran the regression model and generated the output.

The results of the item bias analyses are presented in table 26.2 and an example of PROCESS Macro output is presented in table 26.3. Similar to the ANOVA procedure, we are interested in the main effect of the group (*gender*) and the interaction effect (*gender x total*) to examine uniform and non-uniform bias, respectively. The “model summary” section of the output (see table 26.3) displays these effects. For the item “*careless*”, for example, we observe a significant main effect of gender group ($B = 0.43, t(226) = 3.47, p < .001$) and a non-significant interaction effect ($B = -0.03, t(226) = -1.48, p = .141$). These results indicate the presence of uniform bias for the item “*careless*”.

<TABLE 26.2 AND 26.3 HERE>

In the Johnson-Neyman procedure, conclusions regarding biased items do not have to be based solely on the statistical significance levels. As discussed previously, an advantage of the Johnson-Neyman procedure is that it provides a “region of significance”, which informs us about the amount of sample within the region where the conditional effect of *group* on *item* is significant. The “Johnson-Neyman Technique” section of the output in table 26.3 provides us the Johnson-Neyman point for the moderator (i.e., 5.42). An examination of the regression results in this section reveals that gender group had a significant effect on item score for all cases with a mean centered total score below 5.42. Of particular importance here is not only the Johnson-

Neyman point, but also the amount of sample present within the region of significance. For example, if there are hardly any respondents who have scores below the Johnson-Neyman point, we should be cautious in making item bias claims. The PROCESS Macro output provides this necessary information. As seen on the “Johnson-Neyman Technique” section of the output in table 26.3, about 76.5% of the cases in the data set had values below the Johnson-Neyman point when the item “*careless*” was analyzed. On the other hand, when the item “*disorganized*” was taken into consideration, there was again a significant effect of gender group, indicating presence of uniform bias, but only 27% of the cases were within the significance region (see table 26.2). Researchers should be cautious in making conclusions about item bias in such cases where there are insufficient amount of cases within the region of significance. Although there are no rules of thumb, it is reasonable to expect more than half of the sample to be within the region of significance to have substantial evidence of item bias.

The Johnson-Neyman region of significance can also be plotted graphically for visual inspection of the results. The last section of the output in table 26.3 provides the syntax codes to produce plots in SPSS. These codes can be executed using the SPSS syntax window. Examples of plots for an unbiased item (*practical*) and an item with uniform bias (*careless*) are presented in figure 26.2. Note that the regression lines for groups have a substantial overlap when the item is unbiased (left), but there is a large gap between the regression lines when uniform bias is present (right).

<FIGURE 26.2 HERE>

To test whether the results were stable in a smaller sample, item bias analyses were repeated in a random subsample from the dataset ($N = 100$, 43 Females, 57 Males). The results

were nearly identical (see table 26.4): The significant effect of gender group was again evident for items “*systematic*” and “*careless*”. The only difference was that the significant gender group x total score interaction effect on item “*systematic*” did not emerge in the subsample analysis. This finding was anticipated, given that detecting interactions in multiple regression analysis generally require larger sample sizes (McClelland & Judd, 1993). Therefore, the Johnson-Neyman procedure may not be reliable to detect non-uniform bias when the sample size is small. Small sample sizes (e.g., $N = 100$) are nevertheless not ideal for item-bias analysis in general.

<TABLE 26.4 HERE>

Concluding Remarks

Methods are indispensable to theory development (Greenwald, 2012). Advancing item bias methods and making them available to a wider audience will not only yield more valid group comparisons and better refined measurement instruments, but also enhance theory development through understanding of why specific items are biased across groups. Therefore, as suggested by Zumbo (2007), new generation item bias methods should be made accessible to researchers who are not measurement specialists. The current chapter is an attempt in this direction.

The present research has introduced the Johnson-Neyman procedure as an item bias analysis tool. Future research can advance the ideas presented here through simulation studies where the ANOVA and the Johnson-Neyman procedures are compared with respect to their sensitivity to detect biased items across different sample sizes. This will be a welcome contribution to the literature on item bias analysis.

Acknowledgments

I am grateful to Professor Sigurd V. Troye for his valuable comments to an earlier draft of this chapter.

References

- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*(6), 561-571.
- Caprara, G. V., Barbaranelli, C., Bermúdez, J., Maslach, C., & Ruch, W. (2000). Multivariate methods for the comparison of factor structures in cross-cultural research: An illustration with the Big Five questionnaire. *Journal of Cross-Cultural Psychology*, *31*(4), 437-464.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159.
- D'Alonzo, K. T. (2004). The Johnson-Neyman Procedure as an alternative to ANCOVA. *Western Journal of Nursing Research*, *26*(7), 804-812.
- Fitzsimons, G. J. (2008). Editorial: Death to dichotomizing. *Journal of Consumer Research*, *35*(1), 5-8.
- Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, *7*(2), 99-108.
- Hayes, A. F. (2013). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. New York, NY: Guilford Press.
- Irwin, J. R., & McClelland, G. H. (2003). Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*, *40*(3), 366-371.
- Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, *1*, 57-93.
- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19-40.
- Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*(1), 181-190.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*(2), 376-390.

Miyazaki, Y., & Maier, K. S. (2005). Johnson–Neyman type technique in Hierarchical Linear Models. *Journal of Educational and Behavioral Statistics*, *30*(3), 233-259.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*(4), 437-448.

Ramírez-Esparza, N., Gosling, S. D., Benet-Martínez, V., Potter, J. P., & Pennebaker, J. W. (2006). Do bilinguals have two personalities? A special case of cultural frame switching. *Journal of Research in Personality*, *40*(2), 99-120.

Reynolds, C. R. (2000). Methods for detecting and evaluating cultural bias in neuropsychological tests. In E. Fletcher-Janzen, T. Strickland & C. R. Reynolds (Eds.), *Handbook of Cross-Cultural Neuropsychology* (pp. 249-285). New York: Kluwer Academic/Plenum Publishers.

Santor, D. A., Ramsay, J., & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment*, *6*(3), 255-270.

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment*, *63*(3), 506-516.

Smith, L. L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin*, *28*(6), 754-763.

- Spiller, S. A., Fitzsimons, G. J., Lynch, J. G., & McClelland, G. H. (2013). Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *Journal of Marketing Research*, 50(2), 277-288.
- Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, 99(1), 118-128.
- van de Vijver, F., & Leung, K. (1997). *Methods and Data Analysis for Cross-Cultural Research*. Thousand Oaks, CA: Sage.
- van de Vijver, F., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto & F. van de Vijver (Eds.), *Cross-Cultural Research Methods in Psychology* (pp. 17-45). New York, NY: Cambridge University Press.
- Vecchione, M., Alessandri, G., & Barbaranelli, C. (2012). The Five Factor Model in personnel selection: Measurement equivalence between applicant and non-applicant groups. *Personality and Individual Differences*, 52(4), 503-508.
- West, S. G., Aiken, L. S., & Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality*, 64(1), 1-48.
- Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223-233.

Descriptive Statistics and Independent Samples T-Test Results for the Conscientiousness Scale

	Gender		<i>t</i>	<i>p</i>	Cohen's <i>d</i>
	Females (<i>N</i> = 120)	Males (<i>N</i> = 110)			
Organized	5.23 (1.28)	4.55 (1.41)	3.83	<.001	0.51
Efficient	5.05 (0.96)	4.86 (1.26)	1.26	.206	0.17
Systematic	5.29 (1.16)	4.51 (1.39)	4.65	<.001	0.62
Practical	5.02 (1.14)	4.79 (1.26)	1.42	.157	0.19
Disorganized (R)	5.59 (1.14)	4.90 (1.38)	4.14	<.001	0.55
Sloppy (R)	5.55 (1.08)	5.31 (1.10)	1.68	.094	0.22
Inefficient (R)	5.62 (1.22)	5.31 (1.19)	1.93	.055	0.26
Careless (R)	5.51 (1.28)	5.54 (1.23)	-0.17	.866	0.02
Conscientiousness (Scale)	5.36 (0.78)	4.97 (0.87)	3.54	<.001	0.47
Alpha Coefficient	$\alpha = .83$	$\alpha = .83$			

Note. (R) denotes reverse-coded item.

Table 26.2

Moderated Regression Analysis Results (N = 230)

Items	B	SE B	<i>t</i>	<i>p</i>	Sample within the Johnson-Neyman significance region (%)
Organized					
Gender	-0.20	0.12	-1.67	.096	—
Gender x Total Score	0.01	0.02	0.80	.422	—
Efficient					
Gender	0.12	0.12	0.93	.351	—
Gender x Total Score	0.01	0.02	0.34	.736	—
Systematic					
Gender	-0.31	0.11	-2.85	.005	64.3%
Gender x Total Score	0.03	0.01	2.18	.030	—
Practical					
Gender	0.03	0.15	0.19	.851	—
Gender x Total Score	-0.01	0.02	-0.27	.789	—
Disorganized (R)					
Gender	-0.27	0.13	-2.09	.038	27%
Gender x Total Score	0.01	0.02	0.74	.462	—
Sloppy (R)					
Gender	0.12	0.11	1.09	.278	—
Gender x Total Score	0.01	0.02	0.14	.891	—
Inefficient (R)					
Gender	0.09	0.12	0.71	.478	—
Gender x Total Score	-0.03	0.02	-1.81	.071	—
Careless (R)					
Gender	0.43	0.12	3.47	.001	76.5%
Gender x Total Score	-0.03	0.02	-1.48	.141	—

Note. (R) denotes reverse-coded item.

Table 26.3

Example PROCESS Macro Output for the Johnson-Neyman Analysis of the Item “Careless”

```

***** PROCESS Procedure for SPSS Release 2.13 *****
      Written by Andrew F. Hayes, Ph.D.      www.afhayes.com
      Documentation available in Hayes (2013). www.guilford.com/p/hayes3
*****
Model = 1
      Y = CARELESS
      X = GENDER
      M = TOTAL

Sample size
      230
*****
Outcome: CARELESS

Model Summary
      R      R-sq      MSE      F      df1      df2      p
      ,6889      ,4746      ,8401      58,1149      3,0000      226,0000      ,0000

Model
      coeff      se      t      p      LLCI      ULCI
constant      5,4973      ,0625      87,9862      ,0000      5,3741      5,6204
TOTAL          ,1319      ,0107      12,3828      ,0000      ,1109      ,1529
GENDER         ,4308      ,1241      3,4724      ,0006      ,1863      ,6753
int_1          -,0318      ,0215      -1,4764      ,1412      -,0742      ,0106

Interactions:
int_1  GENDER  X  TOTAL
***** JOHNSON-NEYMAN TECHNIQUE *****
Moderator value(s) defining Johnson-Neyman significance region(s):
      Value      % below      % above
      5,4176      76,5217      23,4783

Conditional effect of X on Y at values of the moderator (M)
      TOTAL      Effect      se      t      p      LLCI      ULCI
-21,3826      1,1107      ,5234      2,1220      ,0349      ,0793      2,1421
-19,5826      1,0535      ,4856      2,1692      ,0311      ,0965      2,0104
-17,7826      ,9962      ,4480      2,2235      ,0272      ,1134      1,8791
-15,9826      ,9390      ,4107      2,2866      ,0231      ,1298      1,7482
-14,1826      ,8818      ,3736      2,3605      ,0191      ,1457      1,6179
-12,3826      ,8245      ,3368      2,4480      ,0151      ,1608      1,4882
-10,5826      ,7673      ,3006      2,5526      ,0114      ,1750      1,3596
-8,7826       ,7101      ,2651      2,6785      ,0079      ,1877      1,2324
-6,9826      ,6528      ,2306      2,8304      ,0051      ,1983      1,1073
-5,1826      ,5956      ,1978      3,0111      ,0029      ,2058      ,9854
-3,3826      ,5384      ,1675      3,2139      ,0015      ,2083      ,8685
-1,5826      ,4811      ,1414      3,4023      ,0008      ,2025      ,7598
,2174       ,4239      ,1222      3,4679      ,0006      ,1830      ,6648
2,0174      ,3667      ,1135      3,2295      ,0014      ,1429      ,5904
3,8174      ,3094      ,1177      2,6298      ,0091      ,0776      ,5413
5,4176      ,2586      ,1312      1,9705      ,0500      ,0000      ,5171
5,6174      ,2522      ,1334      1,8901      ,0600      -,0107      ,5151
7,4174      ,1950      ,1574      1,2389      ,2167      -,1152      ,5051
9,2174      ,1377      ,1864      ,7390      ,4607      -,2295      ,5050
11,0174     ,0805      ,2184      ,3686      ,7128      -,3499      ,5109
12,8174     ,0233      ,2524      ,0922      ,9266      -,4740      ,5206
14,6174     -,0340      ,2875      -,1181      ,9061      -,6006      ,5327
*****
Data for visualizing conditional effect of X on Y
Paste text below into a SPSS syntax window and execute to produce plot.

DATA LIST FREE/GENDER TOTAL CARELESS.
BEGIN DATA.

      -,4783      -6,7721      4,2949
      ,5217      -6,7721      4,9410
      -,4783      ,0000      5,2912
      ,5217      ,0000      5,7220
      -,4783      6,7721      6,2876
      ,5217      6,7721      6,5031

END DATA.
GRAPH/SCATTERPLOT=GENDER WITH CARELESS BY TOTAL.
    
```

Table 26.4
Moderated Regression Analysis Results from a Subsample (N = 100)

Items	B	SE B	t	p	Sample within the Johnson-Neyman significance region (%)
Organized					

Gender	-0.17	0.17	-1.03	.308	
Gender x Total Score	0.03	0.03	1.05	.296	—
Efficient					
Gender	0.29	0.19	1.57	.119	
Gender x Total Score	0.01	0.03	0.14	.889	—
Systematic					
Gender	-0.43	0.16	-2.67	.009	69%
Gender x Total Score	0.02	0.02	1.03	.304	
Practical					
Gender	-0.18	0.20	-0.88	.382	
Gender x Total Score	-0.01	0.03	-0.55	.580	—
Disorganized (R)					
Gender	-0.21	0.19	-1.11	.271	
Gender x Total Score	0.01	0.02	0.47	.641	—
Sloppy (R)					
Gender	0.16	0.17	0.92	.360	
Gender x Total Score	-0.01	0.02	-0.11	.910	—
Inefficient (R)					
Gender	0.02	0.18	0.11	.909	
Gender x Total Score	-0.02	0.02	-0.89	.382	—
Careless (R)					
Gender	0.52	0.18	2.86	.005	68%
Gender x Total Score	-0.03	0.03	-1.01	.312	

Note. (R) denotes reverse-coded item.

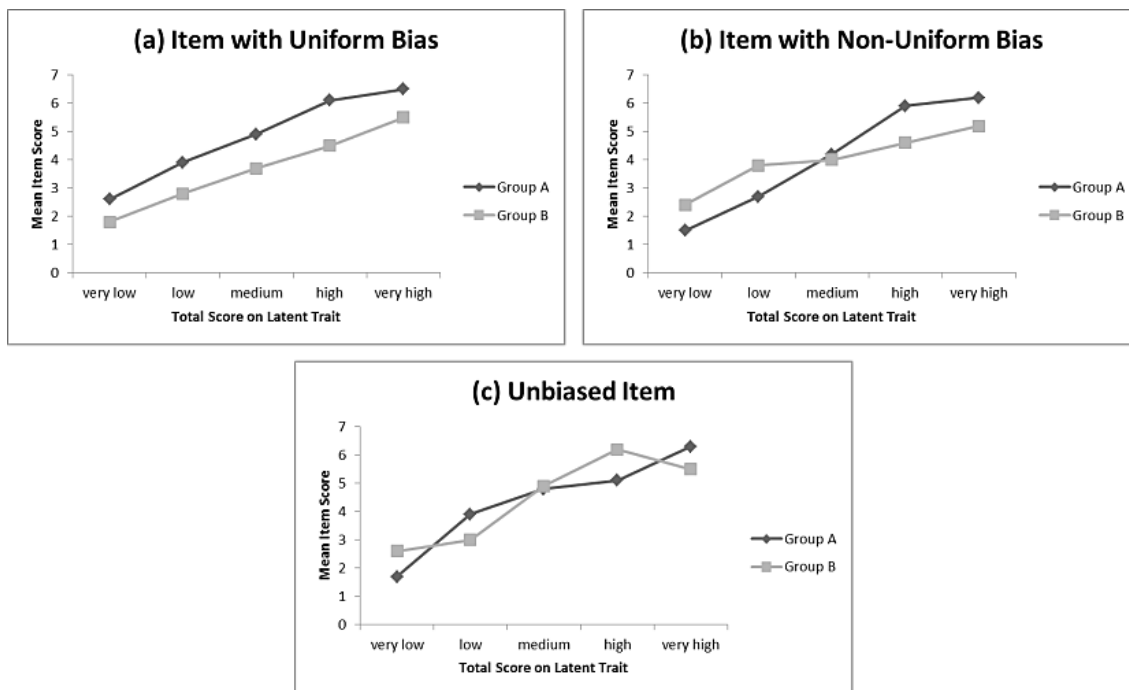


Figure 26.1. Hypothetical examples of an item with (a) uniform bias, (b) non-uniform bias, and (c) no bias (adapted from van Vijver & Leung, 2011).

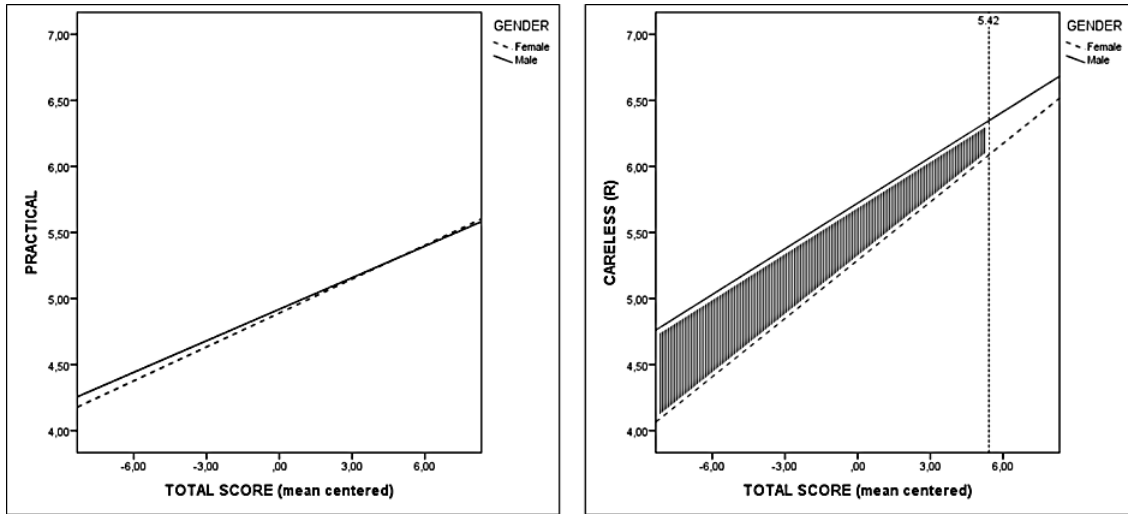


Figure 26.2. Example graphical displays of the Johnson-Neyman analysis results (unbiased item to the left and biased item to the right). The Johnson-Neyman region of significance is shaded in gray. (R) denotes reverse-coded item. Source: Author.