Psihologia Resurselor Umane, **15** (2017), 138–148

http://dx.doi.org/10.24837/pru.2017.2.480

RESEARCH ARTICLE

Assessment of differential item functioning in personnel selection

CRISTIAN OPARIUC-DAN
"Ovidius" University, Constanța, Romania

MIHAELA GRIGORAȘ
University of Bucharest, Romania

ANDREEA BUTUCESCU
University of Bucharest, Romania

ALEXANDRU-ANDREI SÎRBU
University of Bucharest, Romania

ANDREI COSMIN DUMBRAVĂ
University of Bucharest, Romania

Abstract

A psychological test should be developed to accurately estimate the measured construct, being insensitive to extraneous factors that could affect it. The way in which the various external factors jeopardize the item responses is known as test bias and minimizing these influences is a primary concern for test developers. A relatively recent category of psychometric techniques that aim to identify those items that function differently in a group of test-takers is called Differential Item Functioning (DIF). The purpose of this paper is to present, in a non-technical manner, the conceptual foundation of DIF, and the most common applied techniques. We will discuss methods based on logistic regression, those using Item Response Theory models as well as the Mantel-Haenszel method. We will then discuss the strengths and limitations of each method, conditions of application, and the main software packages used to conduct such analyses.

Keywords

Differential item functioning, validity, test bias

Introduction

Additionally to the interview, the abilities and personality tests represent the most widely used selection tools for candidates in employment processes. The use of psychometric tests provides at least theoretically greater rigor in staff decisions and a lower chance of choosing employees incompatible with the workplace. However, using a measure that is "valid and reliable" is not always enough to be useful in organizational practice. A question that every practitioner has to ask is to what extent the sample of participants is similar to the one with which candidates are compared. Most

Correspondence concerning this article should be addressed to Cristian Opariuc-Dan, Administrative Sciences Department, Faculty of Law and Administrative Sciences, University Ovidius, Bd. Mamaia, nr. 124, Aleea Universității nr.1 Constanța, România. Email: copariuc@gmail.com.

evaluation tools are calibrated on a volunteer population, under conditions very different from those in a selection situation. Selection measures are often developed and normalized on samples from the general population in low stake contexts but administered to job applicants for various HR processes in high stake process. There are some challenges to be addressed when considering the use of a measure developed in different conditions from that of selection/screening conditions (De Fruyt, De Bolle, McCrae, Terracciano, and Costa, 2009). For example, in the case of a personality assessment at work, it is related to the falsification and response distortion (Ones, Dilchert, Viswesvaran, and Judge, 2007; Morgeson, Campion, Dipboye, Hollenbeck, Murphy, Schmitt, 2007). In case of abilities tests, groups may differ in their test-taking motivation, and it is a matter of debate whether, apart from mean differences, items or scales function similarly across research and selection/development contexts.

For example, we do not doubt that the Big Five model had five factors confirmed and reconfirmed in the research settings. However, this is not the case of the applicant population. An exploratory factor analysis (Schmitt & Ryan, 1993) revealed that a six-factor solution fits best on the sample of applicants. Schmit and Ryan called this sixth factor an "ideal employee," noting that it „included a conglomerate of item composites from across four of the five subscales of the NEO-FFI" (Schmitt & Ryan, 1993, p. 971). Similar results were also obtained by others researchers like Cellar, Miller, Doverspike, and Klawsky (1996) or Lim and Ployhart (2006). Both the scales (or factor) and the items used in research may function differently in a context with high stakes for individuals. For example, Stark, Chernyshenko, Chan, Lee, and Drasgow (2001) found a differential item and test functioning determined by faking for the Sixteen Personality Factor Questionnaire. Lack of measurement equivalence in data across contexts implies that observed mean differences on relevant constructs might result from measurement artifacts rather than actual differences across individuals. Unless measurement invariance is established, conducting cross-group comparisons of mean differences or other structural parameters is meaningless. The lack of invariance at scale and item level identified in previous research is a significant problem given the widespread use of these types of measures in practice. Making selection decisions on partially known measurement structures are improbable to lead to sound selection decisions (Ion & Iliescu, 2017).

In the process of establishing the validity of psychological diagnostic instruments situations in which item responses are influenced by one or more group variables, such as gender, age category, profession, occupation are common. Indeed, an item like: "When I was a kid I enjoyed playing soccer in the schoolyard", addressed mainly to male respondents, affect the item response because it excludes or severely limits the relevance for female respondents. This, in turn, raises serious questions concerning the validity of measurement inferences. This type of error is called item bias or item level bias and the study of this kind of errors is part of a relatively recent field called test bias analysis. One of the most relevant techniques for which such biases can be studied is called Differential Item Functioning (DIF).

This paper is structured as follows: a first part providing a theoretical and conceptual framework for DIF analyses in the context of current conceptions of validity and fairness, and a second part that reviews the commonly used methods and techniques of DIF analysis, discussing the strengths and limitations of each method, conditions of application as well as software packages implementing them.

Theoretical foundation of DIF

Validity, fairness, item bias and DIF

The traditional view of validity defines it as a property of the measurement instrument, and the process of validation includes demonstrating content, criterion and construct validity. Current view of validity considers it o property of inferences made based on the measurement tool, more precisely a property of the measurement made with the instrument (Zumbo, 1999). The focus is no longer on

statistical procedures, but on consequences of test decisions and use. In this context, to document the interpretation and application of the tests cores becomes important (Bachman, 1990). So if we are to record consequences of test decisions and use, apart from the traditional procedures used to demonstrate the three forms of validity, at present we need to investigate the presence of bias determined by confounding variables.

As such, the concept of validity goes beyond the instrument to include the consequences of test scores use, starting from the definition of the measured construct, continuing with the item analyses and ending with investigating the impact of measurement, namely the consequences determined by the test used for diagnostic or research purposes. Within these analyses test bias and DIF analyses have a particular role.

Confounding variables that are not related to the measured construct, like gender, race, educational level, socio-economic status, shouldn't influence the item responses and, indirectly, the total score. The objective of DIF analyses is that of identifying possibly biased items and explaining the source of these influences because biases can have severe consequences on test scores use, consequences which are difficult to correct via classical methods. The psychometric techniques used in this type of analyses are called methods of fairness analysis, DIF being just a part of this domain. More precisely, DIF is the investigation of *the way in which group membership can influence item performance.*

Differential Item Functioning: definition and types

An item similar to the one presented above ("When I was a kid I enjoyed playing soccer in the schoolyard") could measure the preference for practicing sports. We can presume that persons with the same level of choice for exercising will endorse such an item. If our assumption proves to be false, how can we explain this? We can easily notice the soccer reference, a sport that is preferred to a greater extent by boys rather than girls. A female respondent could not endorse such an item, even if she has the same level of preference for practicing sports. It may be that she enjoys playing basketball, volleyball or aikido and that she may not have a preference for soccer.

In this case, we say that DIF is present, namely that there is a probability that test-takers from different groups with the same level of the latent factor will have different item performances (Clauser & Mazor, 1998). Therefore, when considering DIF, we must take into account three elements: (a) respondents should have the same level of the latent factor, (b) a group variable by which respondents can be divided into independent groups and (c) the latent factor which determines DIF should not be part of the measured construct. If the three conditions are cumulatively met we can consider the presence of DIF (Karami, 2012). We should notice that the presence of DIF, as an effect of an external variable is not necessarily evidence of item bias, the DIF being a source of error only if it has no association with the measured latent factor.

The DIF analysis begins with identifying the variables (preferably categorical variables) that could have an effect on item performance at the same levels of the latent factor. These variables are called **DIF variables**. A DIF variable can affect the item responses in two ways: uniformly and non-uniformly.

Uniform DIF is detected when one group has higher levels of the latent factor in comparison to the other group on the entire continuum of the latent factor (see Figure 1).

Non-uniform DIF is identified when up to a certain level of the latent factor, one group has higher values, and from that point on this tendency disappears to be reverted (see Figure 2).

As we already mentioned, the presence of DIF is not necessarily an indication of a biased item. If the factor that determines the DIF is associated with or relevant for the measured construct, then we refer to item impact rather than item bias. We consider the presence of item bias only if there is no connection between the factor that determines the DIF effect and the measured construct. In the example above, gender, the DIF variable, is in no way associated with the level of exercising,

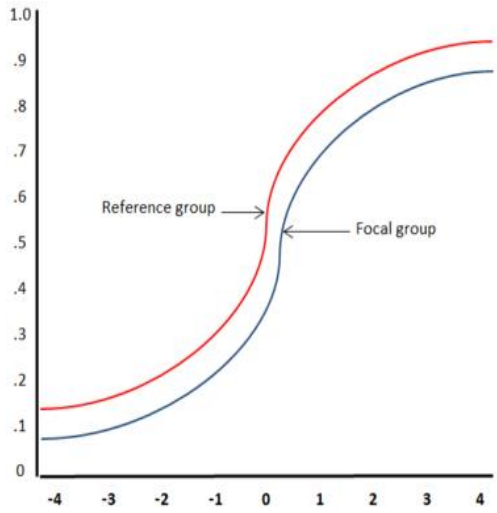

Figure 1. The item characteristic curves of an item displaying uniform DIF (Wikipedia, 2017).

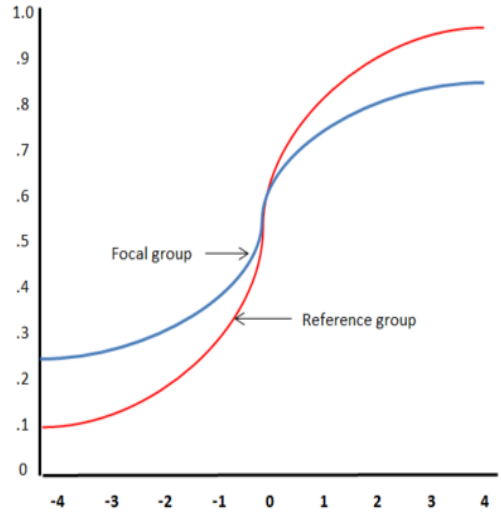

Figure 2. Item characteristic curves of an item displaying non-uniform DIF (Wikipedia, 2017).

which leads us to presume that gender has a biasing influence on the item. In the case of factors related to exercising, such as membership in a martial arts club, we would refer to item impact rather than item bias. However, there are no criteria for deciding the existence of a relation between the DIF variable and the measured construct, the association being estimated in a subjective manner or by a panel of experts.

Finally, we will address the issue of the measured trait. Most often, the studied item is part of a measuring instrument, be it a test or a scale. To be analyzed regarding impact or bias determined by a DIF variable, the item should be related to the measured construct, often the measure being the total score, referred to as **internal criterion**. It is also possible to analyze an item using a different measure of the construct (**external criterion**), but this method implies demonstrating both the reliability of the construct measurement and the fact that the item measures the construct (McNamara, Roever, & Young, 2007).

DIF Methods and Techniques

There are two major methodological approaches to differential item functioning analyses (Magis, Beland, Tuerlinckx, & De Boeck, 2010):

- *Non-IRT methods* — they do not demand the estimation of the parameters of an IRT model and use **the total score as a criterion**. These type of methods use, in general, non-parametric procedures, and are also referred to as non-parametric methods of DIF analyses.
- *IRT methods* — they are based on IRT models, use **the latent factor level as a criterion** and require the estimation of the items' parameters. They are robust, parametric methods, but assumptions and difficult to meet.

DIF analyses can be conducted using IBM SPSS Statistics for instance (Zumbo, 1999), or software packages designed especially for DIF analyses like: „difR" (Magis, Beland, Tuerlinckx, & De Boeck, 2010), lordif"(Choi, Gibbons, and Crane, 2011), „deltaPlotR" (Magis & Facon, 2014).

Mantel-Haenszel (MH) method

The Mantel-Haenszel (MH) method is one of the oldest and most popular non-IRT ways to identify DIF and it relies on *the analysis of the association between the DIF variable and the item response at each level of the total score* (Mantel & Haenszel, 1959). The probabilities of correct and incorrect answers are separately calculated for respondents in the reference group and the focal group. Then, the ratio between these probabilities for each group is computed in order to *determine how likely it is for members of a group to endorse the correct response to an item.* Finally, the two ratios are compared in order to determine *what are the odds for members of the reference group to* endorse the correct item response *compared to the members of the focal group.* Summing the values thus obtained for each level of the total score and dividing them by the number of overall score levels we compute the Mantel-Haenszel odds ratio (α_{MH}), the primary index of this method (Karami, 2012). The Mantel-Haenszel odds ratio (α_{MH}), follows a chi-square distribution with one degree of freedom under the null hypothesis of no association between item responses and group membership, conditionally upon the total score level. If the value of this indicator exceeds the threshold value for a chosen level of significance, then the item will be identified as displaying DIF (Magis, Beland, Tuerlinckx, & De Boeck, 2010).

The logarithm of this indicator (λ_{MH}) follows a normal distribution, values close to zero indicating the absence of DIF, but in practice an adjusted version of this logarithm (Δ_{MH}=-2,35 λ_{MH}) named **ETS Delta** is preferred. Guidelines based on the absolute values of **ETS Delta** have been proposed for the effect size (Holland & Thayer, 1988) which is considered **negligible** for *values smaller or equal to 1,* **moderate** for *values between 1 and 1.5***,** and **large** for *values above 1.5*.

Logistic regression

Logistic regression is one of the most influential methods in DIF analyses, used for the detection of both of uniform and non-uniform DIF. This method is based on the logistic distribution, paving the way for the IRT methods (Swaminathan & Rogers, 1990). Having a number of dichotomous items loading on the same latent factor, item performance (the dichotomous dependent variable) can be predicted by the total score (continuous predictor), the total score and group membership (dummy coded predictor) or the overall score, group membership and the interaction between them. Thus, the statistic model of logistic regression used to detect uniform and non-uniform DIF is:

$$ln_{\left(\frac{P_{active}}{1-P_{active}}\right)} = \beta_0 + \beta_1 Total + \beta_2 Group + \beta_3 Total \times Group + \varepsilon$$

If, using just the total score as a predictor small residuals are obtained, we can state that *the total score alone can predict the probability of a correct answer to the item* and no DIF is detected:

$$ln_{\left(\frac{P_{active}}{1-P_{active}}\right)} = \beta_0 + \beta_1 Total + \varepsilon$$

If, by adding the DIF variable (the grouping variable) the precision of the prediction increases in a statistically significant manner, then we identify an effect of *group membership on the probability of a correct response to the item*, detecting a **uniform DIF effect**.

$$ln_{\left(\frac{P_{active}}{1-P_{active}}\right)} = \beta_0 + \beta_1 Total + \beta_2 Group + \varepsilon$$

Finally, if besides the total score and the grouping variable, the interaction between them significantly increases the predictive power, then we identify an *effect determined both by the total score and group membership on the probability of a correct response to the item*, detecting **non-uniform DIF**.

The analysis is a stepwise one, building, successively, the three models, identifying the effect by either the Wald test or by the probability report test, statistics following a theoretical chi-square distribution with one degree of freedom.

Logistic regression requires a larger number of respodents compared to other methods, a minimum of 200 observations being recommended in each group determined by the DIF variable (Zumbo, 1999).

In „difR" package there are several functions for implementing logistic regression: „*difLogistic*" used for dichotomous items and DIF variables with two categories, *„difGenLogistic"* used for grouping variables with more than two categories, and *„difLogReg"* that allows a continuous group variable (like, for instance, age).

IRT-based methods

So far, the total score was used as an internal criterion for all DIF analyses. For IRT–based methods the total score is replaced by the estimated level of the latent factor for individuals (θ). IRT-based methods are more effective and more informative, but with assumptions more challenging to meet.

Similar to the logistic regression, IRT methods are based on the item characteristic curve (ICC) or item response function, a logistic curve which depicts the relation between latent factor level and the probability of a correct item response (see Figure 3).

According to the ICC, the likelihood of a correct item response to a dichotomous item depends on four parameters:

- **Latent factor coverage level (difficulty) of the item** (parameter **b**) – is the latent factor level *(θ)* that the respondents must possess to have a .5 probability of a correct answer to the item (to have 50% chance that to endorse the right answer).

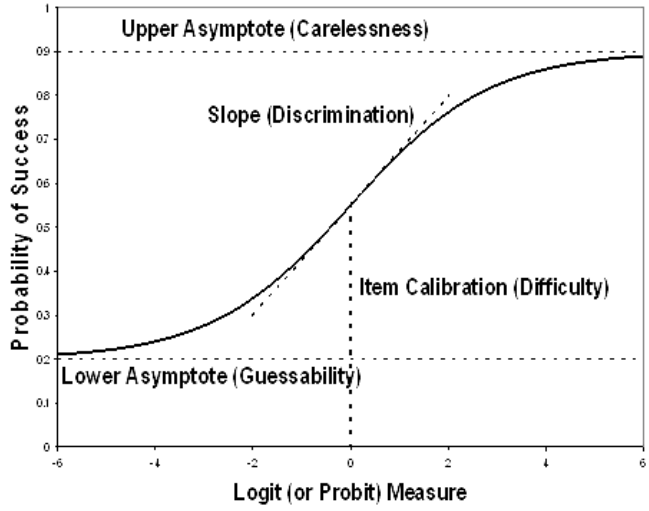

Figure 3. 4 PL Logistic model (complete model). Source: (Item Discrimination, Guessing and Carelessness Asymptotes: Estimating IRT Parameters with Rasch, 2017)

- **Discrimination** (parameter **a**) – *it's given by the slope of the item characteristic curve and represents the increase in probability of a correct answer as the latent factor increases*, or the ability of the item to discriminate between two individuals with very close latent factor levels.

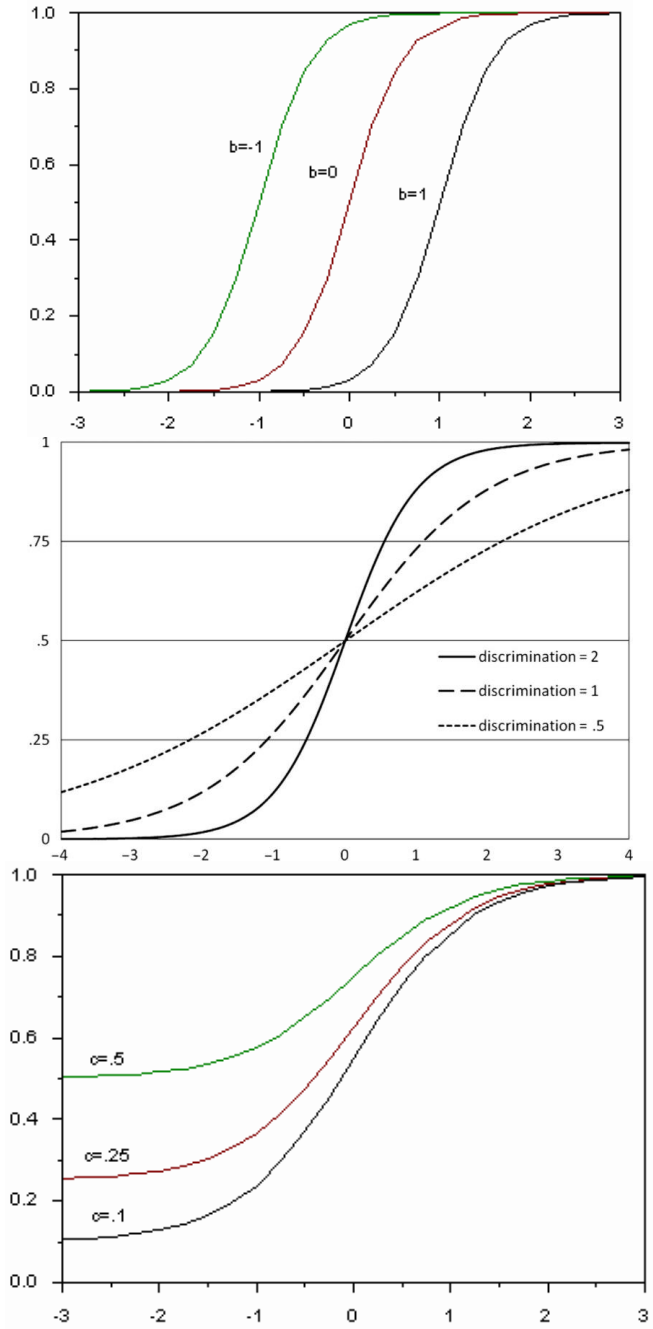

Figure 4. Item characteristic curves showing: (a) differences in item difficulty parameters (top left), (b) differences in item discrimination parameters (top right), (c) differences in guessing parameters (bottom)

- **Guessing** (parameter **c**) – *it's the probability for test-takers with very low levels of the latent factor to endorse the correct item response*, also known as the pseudo-chance parameter.
- **Carelessness** (parameter **d**) – represents the inverse probability of guessing and it refers to the upper asymptote of the item characteristic curve, being *the probability with which the individuals with an extremely high level of latent factor endorse the correct item response*. Typlically, this probability should be close to 1, meaning that the item can be solved, for this reason the four parameter models being rarely used. Low values of this parameter are interpreted rather as neglect, disinterest in finding the correct response or boredom.

Based on these parameters, there are 4 measurement models for dichotomous items:

- **IP**L – a model which uses just the level of coverage in the latent factor – difficulty (parameter b), assuming the same discrimination, guessing and carelessness levels.
- **2PL** – a model which considers both difficulty (parameter b) and discrimination (parameter a), also called Birnbaum model;
- **3PL** – a model which introduces guessing (parameter c) known as the Lord model;
- **4PL** – a model in which all the four parameters are used, also called extended Lord model. This is rarely used, because the interpretation of the higher asymptote regarding disinterest and boredom is questionable.

Using IRT models, DIF analyses may identify uniform and non-uniform effects and, function of the measurement model, can be used for dichotomous or polytomous items. DIF techniques that can be used are the logistic regression, more precisely the test of probability report, the Raju area method or the Lord's chi square test.

Methods used for polytomous items

If for dichotomous items there are several DIF methods available, for polytomous items analyses options are limited to: *the generalized version of the Mantel-Haenszel method* (Agresti, 2002), *the ordinal logistic regression* (Zumbo, 1999) or *the analysis of discriminatory logistics function* (Miller & Spray, 1993).

Ordinal logistic regression method

The ordinal logistic regression has the same principles as the normal logistic regression, with the additional introduction of the parallel regression assumption that is the proportional increase of chances for a correct response across the response categories.

$$ln_{(P_{response \leq k})} = \beta_0 + \beta_1 Total + \varepsilon$$

The first model no longer indicates a fixed, but a cumulative probability, the probability *that the item response is in category „k" or lower, is predicted only by the total score*. This is **model 1** of the ordinal logistic regression, an uncontaminated model, in which the total score is enough to predict the response category.

$$ln_{(P_{response \leq k})} = \beta_0 + \beta_1 Total + \beta_2 Group + \varepsilon$$

If by adding the DIF variable (grouping variable) the precision of the prediction placing the item response in the „k" category or lower increases statistically significant, then we can consider *the effect of group membership on the probability to correctly respond to the item*, detecting **uniform DIF**. This model is known as **model 2** of the ordinal logistic regression, representing a hierarchical superior model.

$$ln_{(P_{response \leq k})} = \beta_0 + \beta_1 Total + \beta_2 Group + \beta_3 Total \times Group + \varepsilon$$

Model 3 of the ordinal logistic regression further includes an interaction effect between the total score and group membership. If besides total score and the grouping variable, the interaction between them significantly increases the prediction power of placing the item response in category „k" or lower, then we identify an *effect determined by both total score and group membership on the probability of a correct item response*, detecting non-uniform DIF, corresponding to the complete statistical model. As for normal logistic regression, testing DIF follows a stepwise procedure, comparing the three models, two by two:

- A statistically significant value when comparing **model 3** with **model 1** for a number of two degrees of freedom represents a **global indicator** of DIF presence.
- A statistically significant value when comparing **model 2** with **model 1** for one degree of freedom represents an indicator of the presence of uniform DIF.
- A statistically significant value when comparing **model 3** with **model 2** for one degree of freedom represents an indicator of **non-uniform** DIF.

DIF detection and DIF magnitude

Chi-square statistic is test of significance used for DIF, but it's sensitivity to the sample size, increases the probability of a type I error. The lack of statistical significance indeed indicates no effect, but a statistically significant value does not necessarily indicate the presence of the impact, different magnitude measures being proposed.

For **dichotomous items**, where *the dichotomy does not imply a relationship of order* (where value 1 does not mean „more" compared to 0), two effect size measures have been proposed: **R^2 Nagelkerke** and **R^2 WLS** (weighted-least-squares):

- **R^2 Nagelkerke** (Nagelkerke, 1991) is easy to compute and is included in most data analyses software packages, including IBM SPSS Statistics, but it can only be used with hierarchical regression models, such as those outlined above.
- **R^2 WLS** (Thomas & Zumbo, 1998) uses the weighted version of the smallest squares method and it has the advantage that it doesn't necessarily require hierarchical regression models, could be used also with a simple block design on , making it more robust in terms of assumptions.

The magnitude measure used for ordered ordinal or dichotomous items is **R^2** statistics **for ordinal variables** (McKelvey & Zavoina, 1975), the only statistic that can assume the existence of requested response categories. Therefore, irrespective of the item nature, a thorough analysis will consider both the chi-square test and the effect size measure. Moreover, to identify a general DIF effect, using both **statistical thresholds of p<.01**and **R^2>.13** (Zumbo, 1999) is recommended**,** as we are simultaneously testing several assumptions. The same author (Zumbo, 1999) proposes guidelines for the DIF effect size, implemented in„difR" package (Thomas & Zumbo, 1998): for values of R^2 smaller than .13, the general DIF effect is considered negligible, values between .13 and .29 indicate a moderate DIF effect, and values above .29

indicate a substantial DIF effect. Other researchers considered these thresholds too large, inflating type II error (the risk to consider an item with DIF as unbiased). Lower thresholds had been proposed (Jodoin & Gierl, 2001): values of R^2 smaller than .035 indicating negligible DIF, values between .035 and .07 indicate a moderate DIF, and values above .07 indicate a large DIF effect. Both sets of guidelines were included in the „difR" package.

Analyzing the evolution of the hierarchical model coefficients for uniform DIF, a different criterion for detection of **uniform DIF** was proposed (Crane, Gibbons, Jolley, & van Belle, 2006): the relative difference between the β_1 coefficients of the second and the first model, also known as $\Delta\beta_1$ criterion. For values higher than 10% of $\Delta\beta_1$, uniform DIF is detected. As this value was considered too large, the authors reduced the limit one year later to 5% ($\Delta\beta_1$>.05) or even 1% ($\Delta\beta_1$>.01) to avoid type II error.

„lordif" Package (Choi, Gibbons, & Crane, Package 'lordif': Logistic Ordinal Regression Differential Item Functioning using IRT, 2016) runs DIF analysis on dichotomous and polytomous items using a combination of ordinal logistic regression and the level of the latent factor as an internal criterion, being a hybrid method. In the same time, the package also includes Monte Carlo simulation procedures, which can significantly improve the accuracy of the statistic results, but also to automatically purification of the items.

Conclusion

DIF methods and techniques are valuable tools in I/O psychology providing the means to identify construct irrelevant variance of that measurement tools that could affect the validity of inferences made. While traditionally the CFA approach to establish measurement invariance has been more popular in I/O psychology, currently DIF analysis represent a growing area due to the complementary nature of information they provide. Our paper intended to provide a review of commonly used DIF methods and techniques to stimulate their use in I/O practice.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley-Interscience.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Cellar, D. F., Miller, M. L., Doverspike, D. D., & Klawsky, J. D. (1996). Comparison of factor structures and criterion-related validity coefficients for two measures of personality based on the five factor model. *Journal of Applied Psychology, 81*(6), 694–704.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software, 39*(8), 1–30.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2016, 03 03). *Package 'lordif': Logistic Ordinal Regression Differential Item Functioning using IRT*. Retrieved from https://cran.r-project.org/web/packages/lordif/lordif.pdf

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurements: Issues and Practice, 17*(1), 31–-44.

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care, 44*(11), 115–123.

De Fruyt F., De Bolle M., McCrae R.R., Terracciano A., Costa P.T., Jr. Assessing the universal structure of personality in early adolescence: The NEO-PI-R and NEO-PI-3 in 24 cultures (2009). *Assessment, 6*(3), 301–311.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In P. W. Holland, & H. I. Braun, *Test validity* (pp. 129–145). Hillsdale: Erlbaum.

Ion, A. & Iliescu, D. (2017). Measurement equivalence of personality measures in low-and high-stake testing contexts. *Journal of Personality and Individual Differences, 110*, 1–6.

Item Discrimination, Guessing and Carelessness Asymptotes: Estimating IRT Parameters with Rasch. (2017, October 11). Retrived from https://www.rasch.org/rmt/rmt181b.htm

Jodoin, M. G., & Gierl, M. J. (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*(4), 329–349.

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment, 11*(2), 59–76.

Lim, B., & Ployhart, R. E. (2006). Assessing the convergent and discriminant validity of Goldberg's International Personality Item Pool: A multitrait-

multimethod examination. *Organizational Research Methods*, *9*, 29–54.

Magis, D., & Facon, B. (2014). deltaPlotR: An R package for differential item functioning analysis with Angoff's Delta Plot. *Journal of Statistical Software, 59*(1), 1–19.

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719–748.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology, 4*(1), 103–120.

McNamara, T., Roever, C., & Young, R. F. (2007). *Language Testing: The Social Dimension.* Oxford: Blackwell Publishing.

Miller, T. R., & Spray, J. A. (1993). Logistic Discriminant Function Analysis for DIF Identification of Polytomously Scored Items. *Journal of Educational Measurement, 30*(2), 107–122.

Morgeson, F. P., Campion, M.A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., Schmitt, N. (2007). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60*, 1029–1049.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691–692.

Ones, D. S., Dilchert, S., Viswesvaran, C., Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995–1027.

Stark, S., Chernyshenko, O. S., Chan, K. Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86*(5), 943–953.

Schmit, M.J., & Ryan, A.M. (1993) The big five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology*, *78*, 966–974.

Swaminathan, H., & Rogers, J. H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370.

Thomas, R. D., & Zumbo, B. D. (1998). *Variable importance in logistic regression based on partitioning an R-squared measure*. Presented at the Psychometric Society Meetings, Urbana, Illinois.

Wikipedia (2017, September 25). *Differential item functioning.* Retrieved from https://en.wikipedia.org/wiki/Differential_item_functioning

Zumbo, B. D. (1999). *A Handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-Type (ordinal) item scores.* Ottawa, Ontario: Directorate of Human Resources Research and Evaluation, Department of National Defense.