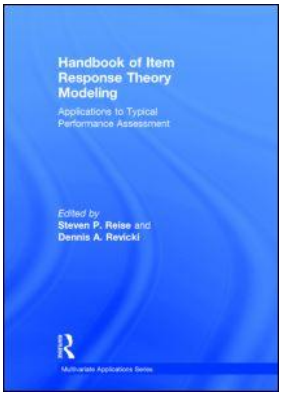
This article was downloaded by: *10.3.97.143*
On: 06 Dec 2023
Access details: *subscription number*
Publisher: *Routledge*
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: 5 Howick Place, London SW1P 1WG, UK



Handbook of Item Response Theory Modeling Applications to Typical Performance Assessment

Steven P. Reise, Dennis A. Revicki

Assessing Person Fit in Typical-Response Measures

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315736013.ch7
Pere J. Ferrando
Published online on: 16 Dec 2014

How to cite :- Pere J. Ferrando. 16 Dec 2014, *Assessing Person Fit in Typical-Response Measures from:* Handbook of Item Response Theory Modeling, Applications to Typical Performance Assessment Routledge
Accessed on: 06 Dec 2023
https://www.routledgehandbooks.com/doi/10.4324/9781315736013.ch7

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

7 Assessing Person Fit in Typical-Response Measures

Pere J. Ferrando

Introduction

The fit of an item response theory (IRT) model to the data is usually assessed by considering the entire sample of test respondents (Chapter 6). Overall model-data fit is assessed by jointly considering all the items × individuals responses. At a more specific level, item fit is assessed on an item-by-item basis by considering the item responses across the group of respondents.

Model-data fit can also be assessed at the level of each individual respondent (person fit) by considering the responses of the individual across the set of test items. Because each individual response pattern contributes to the overall fit of the model (e.g., Reise & Widaman, 1999), overall fit and person fit are necessarily related, and overall fit must be assessed before person fit. A reasonably good overall model-data fit is essential if the IRT model is to be regarded as appropriate. However, an acceptable fit is still compatible with a certain proportion of individuals whose response patterns cannot be adequately explained by the model (Levine & Drasgow, 1983). These patterns will be referred to as misfitting or inconsistent.

Assessing person fit is important for at least three reasons. First, as mentioned earlier, the existence of misfitting patterns can affect the overall fit of the model. And even if this fit is found to be acceptable, their presence might still result in biased estimates of some model parameters (Bollen & Arminger, 1991; Nering, 1997). Second, in validity assessment, scores based on inconsistent patterns could affect the estimated relations between trait levels and relevant external variables (Schmitt, Chan, Sacco, McFarland, & Jennings, 1999). The third and main reason, however, is that if a response pattern is not well explained by the model, there is no guarantee that the score assigned to this pattern adequately reflects the "true" trait level of the individual. If it does not, the invalid score can lead to erroneous decisions. The examples included in this chapter illustrate the importance of this problem in practical settings.

Consider an employment selection process that is (partly) based on a measure of emotional stability. Suppose next that an applicant gives honest answers to the most "neutral" items but deliberately distorts his responses to the most socially desirable items in order to appear more stable than he really is (i.e., faking good). As a result his trait estimate is upwardly biased. Finally, assume that the "true" trait level of this respondent would have placed him below the cutoff value but that the biased estimate places him above, so this applicant is hired. An erroneous decision has been made and the wrong person has probably been selected.

As a second example, consider a clinical scenario in which dysfunctional impulsivity is assessed by means of a test. Consider now: (a) a respondent who is not interested in the assessment and who answers many of the items randomly, and (b) a respondent who

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

answers honestly but who tends to make a disproportionate use of the scale endpoints in many of the items (i.e., an extreme respondent). In both cases the trait estimate of the respondent is likely to be biased, and in case (a) it is probably meaningless. Now, if these estimates were interpreted as if they were valid trait indicators, the assessment of these individuals would probably be highly distorted.

In the examples just discussed, the response pattern of the individuals is expected to be inconsistent to some extent. In the first example, the respondent would give faked responses to some items and honest responses to the rest. In the second example, the respondent would answer some of the items consistently and others at random. In the final example, the respondent would give responses too extreme for her real trait level in some of the items. Now, if these inconsistencies could be detected by using person fit analysis, the flawed interpretations and wrong decisions caused by the blind use of the trait estimates might be avoided.

"Person fit" is a general term that includes all the procedures for assessing response inconsistency at the individual level (Meijer & Sijtsma, 2001). This chapter, however, takes a narrower view, and focuses on procedures that assume a particular parametric IRT model to fit the data. These procedures were initially developed within the maximum performance domain (Reise & Flannery, 1996).

Assessing Person Fit in Typical-Response Measures

Typical-response-based person fit assessment differs from maximum-performance-based assessment in several aspects. I shall discuss this issue in relation to three main points: (a) theoretical relevance, (b) sources of misfit, and (c) psychometric properties.

Theoretical relevance. Person fit procedures in maximum-performance tests have largely been developed for practical purposes, particularly to identify protocols that are invalid because of cheating, guessing, or for other reasons (Meijer, 1996). In typical-response measurement, on the other hand, inconsistent responding has often been linked to theory (Reise & Waller, 1993). In the personality domain in particular, there has been a rich debate on the meaning of intra-individual consistency (e.g., trait relevance or trait organization).

Sources of misfit. The sources of person misfit in typical-response measurement are different from those in maximum-performance measurement. The main sources described so far are: (a) idiosyncratic interpretation of the item content (including problems of understanding item meaning), (b) unmotivated or unsympathetic test responding, (c) multidimensionality, (d) person unreliability/untraitedness, (e) response biases, mainly acquiescence and faking/socially desirable responding, and (f) idiosyncratic response scale usage, including extreme and middle responding (Ferrando, 2010; Meijer, Egberink, Emons, & Sijtsma, 2008; Reise & Flannery, 1996; Reise & Waller, 1993; Waller & Reise, 1992; Zickar & Drasgow, 1996). These sources will be further discussed with regard to (a) the type of inconsistency that they are expected to produce, and (b) the potential effectiveness of the proposed procedures for detecting it.

Psychometric properties. In general, the psychometric requisites that are conducive to valid and statistically powerful person fit assessment are more difficult to attain in the case of personality and attitude measures than in the case of maximum-performance measures. This point is further discussed later in this chapter.

So far, person fit applications based on personality and attitude tests are scarce. There are early applications with an important methodological component, for example: Ferrando and Chico (2001), Reise (1995), Reise and Flannery (1996), Reise and Waller (1993), and Zickar and Drasgow (1996). At the purely applied level, however, it appears

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

that this methodology has started to permeate the field only recently (Conrad et al., 2010; Dodeen & Darabi, 2009; Egberink & Meijer, 2011; Ferrando, 2012; Meijer et al., 2008; Woods, Oltmanns, & Turkheimer, 2008). This state of affairs might be partly due to the fact that most applied researchers remain either unaware or unconvinced of the value of person fit assessment. Person fit research that has been carried out to date has been far too technical and has focused mostly on statistical issues (Meijer, 2003; Meijer et al., 2008). Moreover, as we shall see, person fit procedures still have important practical limitations.

Emons, Sijtsma, and Meijer (2004, 2005) proposed that additional information should be obtained about (a) the type of inconsistency, (b) the type of item in which inconsistency occurs, and (c) the impact that inconsistency has on the trait estimates so that the assessment of individual misfit could be improved. Their proposal is based on a combined use of global scalar indices, graphical procedures, and indices at the level of items or subsets of items. In this chapter I shall propose a closely related approach.

The remainder of this chapter is organized as follows. First, the models on which the person fit procedures proposed here are based are reviewed. Second, the proposed global, graphical, and single-response procedures are discussed. Third, a series of examples based on real-data applications are presented and discussed. Finally, a general discussion is provided.

Review of the Models and Needed Results

The models reviewed in this section are dominance models. So, the expected item score (in the appropriate direction) is assumed to increase with trait level. They are intended for binary, graded, and (approximately) continuous item response formats. Binary items are still quite common in personality assessment (Reise, Waller, & Comrey, 2000). Graded-response items, particularly Likert scales, are the most commonly used in both personality and attitude tests. Finally, graded responses with a large number of points or continuous-limited formats are increasingly being used in computerized administration (Ferrando, 2002).

Consider a typical-response test, made up of n items, that aims to measure a trait θ and that is administered to a respondent i. Let X_{ij} be the observed score of respondent i on item j and assume that θ is scaled in a z-score metric (mean 0 and variance 1) in the population of respondents.

Assume first that the n items use a binary response format (0 or 1). The probability that respondent i will endorse item j is assumed to be given by:

$$P_j(\theta_i) = P(X_{ij} = 1 \mid \theta_i) = \frac{\exp\left[a_j(\theta_i - b_j)\right]}{1 + \exp\left[a_j(\theta_i - b_j)\right]} = \Psi\left(a_j(\theta_i - b_j)\right). \tag{7.1}$$

Equation (7.1) is the logistic version of the two-parameter model (2PM). The location parameter b_j ("item difficulty" in ability measurement) indicates the trait level that is required to have a probability of 0.50 of endorsing this item. The discrimination parameter a_j indicates the quality of the item as a measure of the trait (Lord & Novick, 1968). The higher a_j is, the more precise the item and the more information it provides about the trait that is measured. Finally, the probability of item endorsement when viewed as a function of θ (i.e., $P_j(\theta)$), which is also the regression of the item scores on θ, is called the item characteristic function (ICF) of item j.

When a_j is set at the same value for all the items, the 2PM reduces to the one-parameter model (1PM). In some clinical instruments that measure narrow traits, substantial variations in item discriminating power have been observed (Reise & Waller, 2009). So in this case the 1PM would clearly be inappropriate. However, in many normal-range

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

typical-response items that measure broad traits the variation in discriminating power is relatively modest (Ferrando, 2004; Hulin, Drasgow, & Parsons, 1983; Levy, 1973). If this is so, the 1PM is a model that should be considered.

Assume now that the items use a graded response format with $m + 1$ categories. In this case, item j has m fixed ordered locations or thresholds $b_{j1}<b_{j2}<b_{jr}\ldots<b_{jm}$. As a function of θ, the probability of endorsing category r, which is called the item category response function, is now:

$$P(X_{ij} = r \mid \theta) = \Psi(Da_j(\theta - b_{jr-1})) - \Psi(Da_j(\theta - b_{jr}))\,. \tag{7.2}$$

Equation (7.2) is the logistic version of Samejima's (1969) graded-response model (GRM). And the expected item score that corresponds to a given trait level—in other words, the regression of the item scores on θ—is given by (Chang & Mazzeo, 1994):

$$E(X_j \mid \theta) = \sum_r rP(X_j = r \mid \theta)\,. \tag{7.3}$$

Finally, assume that the items use a continuous response format. The conditional distribution of the item score for fixed θ is now assumed to be normal, with mean and variance given by:

$$E(X_j \mid \theta) = \mu_j + \lambda_j\theta_i \quad ; \quad Var(X_j \mid \theta) = \sigma^2_{\varepsilon j}\,, \tag{7.4}$$

where μ_j is the item intercept, λ_j the item loading, slope, or regression weight, and $\sigma^2{}_{\varepsilon j}$ the variance of the measurement error. The conditional mean in (7.4) is the linear ICF of the model. Model (7.4) is Spearman's Factor Analysis (FA) model, which in the psychometric literature is also known as the congeneric test (item) score model (Jöreskog, 1971). Ferrando (2009) proposed a re-expression of model (7.4) to make it closer to the 2PM in (7.1). To make the comparison even closer, the item scores can be rescaled to have values between zero and one so that 0.5 corresponds to the midpoint of the item response scale. By making the transformation:

$$\beta_j = \frac{1 - 2\mu_j}{2\lambda_j}\,, \tag{7.5}$$

the conditional expectation in (7.4) can be written as:

$$E(X_j \mid \theta) = 0.5 + \lambda_j(\theta - \beta_j). \tag{7.6}$$

The item parameter β_j is now defined on the same scale as θ and is a location index. It can be interpreted as the trait level that corresponds to an expected score of 0.5 (i.e., the response scale midpoint).The slope λ_j is interpreted as the item discrimination index (Mellenbergh, 1994).

Linear FA is by far the most used model for fitting graded or continuous typical-response items (Hofstee, Ten Berge, & Hendricks, 1998). In principle, it is a model intended for continuous-unlimited scores and, because item responses are bounded and to a greater or lesser extent discrete, it can only be approximately correct. As an approximation, however, both theoretical (Culpepper, 2013; Lord & Novick, 1968) and empirical (Muthén & Kaplan, 1985; Olsson, 1979) evidence suggests that the linear model works well with graded or more continuous items when (a) the discriminating power of the items is moderate or low, and (b) the items have no extreme locations. Typical-response items, which measure broad normal-range traits, tend to meet these conditions (Ferrando, 2004;

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

Levy, 1973). On the other hand, clinical items that measure narrow traits sometimes have both high discriminations and extreme distributions (Reise & Waller, 2009). The linear model is clearly inappropriate in this case.

It is assumed that all the models discussed so far are fitted by using a two-stage approach (i.e., calibration and scoring; see McDonald, 1999). During the calibration stage, the appropriateness of the model is first assessed by conducting an overall model-data fit investigation, and if it is judged to be appropriate, the item parameters are estimated. During the scoring stage, the item estimates obtained during the previous stage are taken as fixed and known and used to estimate the individual trait levels. Within this framework, the procedures considered in this chapter assess the extent to which a response pattern is consistent with the pattern that would be expected given (a) the item parameter estimates obtained during the calibration stage, and (b) the trait level estimate of the respondent obtained during the scoring stage.

Many scoring results that are now provided are directly linked to the procedures described in the following sections. The likelihood of a response vector $\mathbf{x}_i$ for each of the models is:

$$\text{1PM and 2PM: } L(\mathbf{x}_i \mid \theta) = \prod_j P_j(\theta)^{X_{ij}} (1 - P_j(\theta))^{1-X_{ij}}. \tag{7.7}$$

$$\text{GRM: } L(\mathbf{x}_i \mid \theta) = \prod_j \prod_r P_{jr}(\theta)^{u_{ijr}}, \tag{7.8}$$

where $u_{ijr} = 1$ if respondent i chooses category r for item j, and $u_{ijr} = 0$ otherwise. And:

$$\text{Congeneric model: } L(\mathbf{x}_i \mid \theta) = \prod_j \left[\frac{1}{\sigma_{\varepsilon j}\sqrt{2\pi}} \exp - \frac{1}{2}\left(\frac{x_{ij} - \mu_j - \lambda_j \theta}{\sigma_{\varepsilon j}} \right)^2 \right]. \tag{7.9}$$

Maximum likelihood (ML) estimates of the trait level of individual i in the three types of models are the values that maximize (7.7), (7.8), and (7.9). In the binary and graded response cases, these estimates must be obtained iteratively. In the congeneric model they can be obtained in closed form and are the well-known Bartlett's factor scores (e.g., McDonald, 1999; Mellenbergh, 1994).

$$\hat{\theta}_i = \frac{\sum_j \frac{\lambda_j (x_{ij} - \mu_j)}{\sigma^2_{\varepsilon j}}}{\sum_j \frac{\lambda^2{}_j}{\sigma^2_{\varepsilon j}}}. \tag{7.10}$$

Research Methods

Current Person Fit Methods

Global Person Fit Indices

Global indices will be discussed by making a distinction between practical and specific indices. From a hypothesis-testing point of view, practical indices test the null hypothesis of consistency against no specific alternative. So, even if the index is capable of detecting that a pattern is inconsistent, it provides no further information regarding the type of inconsistency. In contrast, specific indices test against specific types of misfit (e.g., faking or extreme responding) and so, in principle, they are more powerful and provide more information.

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

Practical Indices

Likelihood-Based Indices

Of the considerable number of practical indices (Karabatsos, 2003; Meijer & Sijtsma, 2001), some of the most popular are still the likelihood-based (L-B) indices initially proposed by Levine and Rubin (1979). They have a clear rationale, are easy to compute, and, although they have limitations, they generally perform equal to or better than alternative indices (Armstrong, Stoumbos, Kung, & Shi, 2007; Drasgow, Levine, & McLaughlin, 1987; Li & Olejnik, 1997; Meijer, 1996; Nering, 1997; Nering & Meijer, 1998; Reise & Due, 1991).

The basic rationale on which L-B indices are based is that that the likelihood function value of a particular item response pattern will be large for patterns that are consistent with the model and small for inconsistent patterns. Assuming that the item parameters are fixed and known, the unstandardized log-likelihood index l_0 is simply the logarithm of the likelihood function evaluated at the maximizing value of θ (i.e., the ML trait level estimate). According to (7.7) and (7.8), the indices corresponding to the binary and the graded response cases for respondent *i*, item *j*, and response category *r* are:

$$l_0(\theta_i) = \sum_j \{X_{ij} \ln P_j(\theta_i) + \left[(1 - X_j)\ln(1 - P_j(\theta_i))\right]\}. \tag{7.11}$$

And:

$$l^{(p)}{}_0(\theta_i) = \sum_j \sum_r u_{ijr} \ln P_{jr}(\theta_i). \tag{7.12}$$

Ideally, a person fit index should: (a) have reference values so that it can be interpreted, (b) be independent of test length, and (c) be independent of the trait level, and so detect misfitting patterns equally well at all levels (Drasgow, Levine, & Williams, 1985). However, l_o and $l_o^{(p)}$ do not comply with any of these requirements. To improve these limitations, Drasgow and colleagues (1985) derived the standardized *lz* versions:

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{Var(l_0)}};\ \ l_z^{(p)} = \frac{l^{(p)}{}_0 - E(l^{(p)}{}_0)}{\sqrt{Var(l^{(p)}{}_0)}}. \tag{7.13}$$

The formulas and rationale for the expectations and variances in (7.13) can be found in Drasgow and colleagues (1985) and Hulin, Drasgow, and Parsons (1983). If the "true" trait levels were known, then the standardized indices in (7.13) would be expected to asymptotically follow a standard normal distribution under the null hypothesis of consistency (Drasgow et al., 1985).

Ferrando (2007) derived two L-B global indices for the congeneric model. According to (7.9), the log-likelihood corresponding to this model is:

$$\ln L(\mathbf{x}_i \mid \theta_i) = \sum_j (\ln \frac{1}{\sigma_{\varepsilon j}\sqrt{2\pi}}) - \frac{1}{2}\sum_j z_{ij}^2, \tag{7.14}$$

where:

$$z_{ij} = \frac{X_{ij} - \mu_j - \lambda_j \theta_i}{\sigma_{\varepsilon j}}. \tag{7.15}$$

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

The first index proposed by Ferrando is:

$$lco(i) = -2(\ln(\mathbf{x}_i \mid \theta_i) - \sum_{j}^{n}(\ln\frac{1}{\sigma_j\sqrt{2\pi}})) = \sum_{j}^{n} z_{ij}^2(\theta_i)\,, \tag{7.16}$$

where $\hat{\theta}_i$ is Bartlett's ML estimate (7.10). Ferrando (2007) showed that, under the model's assumptions the distribution of the individual *lco* values across respondents was χ^2 with $n - 1$ degrees of freedom.

The second index is a normal approximation computed as:

$$lcz_i = \sqrt{2lco_i} - \sqrt{2n-3}\,. \tag{7.17}$$

Conceptually, all the indices discussed so far essentially measure a type of misfit that can be named "violation to a Guttman pattern" (Armstrong et al., 2007; Meijer, 2003). Thus, in the binary case, which is the clearest, a well-fitting, scalable pattern is one in which the respondent tends to endorse the items whose difficulty index is below his/her estimated trait level but not endorse the items whose difficulty index is above this level. Indeed, at the extreme of this trend, the best-fitting patterns as measured by *lz* are those of a Guttman scale. Non-fitting patterns will be those in which the pattern of endorsement is not consistent with the ordering of items by their difficulty. As for interpretation, *lz* and $lz^{(p)}$ are interpreted as a standard normal *z*-score. A large negative value is an indicator that the pattern is inconsistent given the model and the estimated trait value. A large positive value indicates that the pattern is more deterministic than the stochastic model predicts. The *lcz* index is minimum chi-square, and functions in the opposite direction. So large positive values are indicators of misfit.

Limitations of the L-B Indices

The standard-normal reference distribution for *lz* and $lz^{(p)}$ is an asymptotic result obtained by assuming that the "true" trait level is known. Drasgow and colleagues (1985) found that using the true trait levels in tests of 80 or more items produced close agreements with the normal distribution. In practice, however, the true trait level is unknown and an estimate is used in its place. Furthermore, typical-response measures usually have considerably fewer than 80 items. Research suggests that the use of a trait estimate instead of the true level generally leads to a negatively skewed distribution of the statistic, the variance of which is smaller than expected if θ were known (Magis, Raîche, & Béland, 2012; Molenaar & Hoijtink, 1990; Nering, 1995, 1997; Reise, 1995). This second result leads to underdetection of the inconsistent patterns (van Krimpen-Stoop & Meijer, 2002). As expected, the shorter the test is, the more serious this problem becomes.

Corrections for improving these problems have been proposed for *lz*. Snijders (2001; see also Magis et al., 2012) studied the distribution of *lz* when the true θ is replaced by an estimate and proposed a corrected version of the index that asymptotically approaches the standard normal distribution as long as the trait estimate fulfills some restrictions. On the other hand, de la Torre and Deng (2008) proposed a method that (a) uses an improved expected a posteriori (EAP) trait estimate that is corrected for unreliability, and (b) constructs the distribution of the person fit statistic by using resampling methods. This second proposal can be easily extended to $lz^{(p)}$.

The *lcz* index is expected to behave better than *lz* and $lz^{(p)}$ as far as the two limitations discussed earlier are concerned. First, it explicitly takes into account that the ML trait estimate is used instead of the "true" trait level. Second, the distribution is not asymptotic but exact and so it is correct for any test length. However, the index is based on assumptions that can

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

only be approximately correct (linearity, homoscedasticity, and conditional normality). So the χ^2 reference distribution must be considered as an approximation. Further research is clearly needed on the behavior of the index and potential improvements (Clark, 2010).

We turn now to the more general limitations. First, although the item parameters are taken as fixed and known, they are generally estimated in a sample that probably contains an unknown proportion of inconsistent respondents. Second, the same pattern is used to estimate the trait level and to compute the person fit index (e.g., Karabatsos, 2003). As for the first point, the presence of inconsistent respondents in the calibration sample is expected to affect the quality of item parameter estimates and this result, in turn, is expected to lead to poorly estimated or biased individual trait estimates (Nering, 1997). As for the second, when the estimate is used instead of θ, the result is a decrease in detection power, which, to a large extent, comes from the shift in the estimate due to the inconsistent responses (Armstrong et al., 2007). For the first problem, Nering (1997) suggested using a series of successive recalibrations in which the inconsistent patterns were removed from the sample in each step. As for the second, a potential solution is to use improved estimates that minimize the shift noted earlier. Estimates such as expected a posteriori (EAP) that use more information (in the form of a prior), or robust procedures such as the biweight that down-weights the most inconsistent scores are sound candidates (Meijer & Nering, 1997; Reise, 1995).

By far the most important practical shortcoming of unspecific indices is their low detection power. Research (Ferrando, 2004; Molenaar & Hoijtink, 1990; Reise & Due, 1991) clearly shows that the power of L-B indices (and many other person fit indices) depends on three main factors: (a) test length, (b) spread of item locations, and (c) amount of item discrimination. Many typical-response scales are short, too short in fact even to accurately estimate trait levels (Emons, Sijtsma, & Meijer, 2007). They are also made up of items with moderate or low discriminations and developed with little concern for the range of θ at which the test measures accurately. My opinion is that the quality of most typical-response measures needs to be improved. However, accepting the situation for what it is, a direct potential improvement can be made to point (a). Many personality and attitude measures are multidimensional tests made up of several short scales (Ferrando, 2009; Hulin, Drasgow, & Parsons, 1983; Reise & Flannery, 1996). So, provided that inconsistency generalizes over scales, the development of multidimensional indices that are based on all the items in the test might increase the power.

Parsons (see Hulin, Drasgow, & Parsons, 1983) proposed a first heuristic multidimensional extension for discrete-response models. If the dimensions measured by the questionnaire are highly correlated, then lz or $lz^{(p)}$ can be obtained from all of the items as if they formed a common scale. If this is not the case, a multidimensional lz index can be obtained as a weighted average of the unidimensional lz's, the weights being proportional to the number of items in the sub-scale.

Drasgow, Levine, and McLaughlin (1991) proposed a more rigorous extension of this type intended for what they termed a "multi-unidimensional" test (i.e., a test consisting of a series of unidimensional scales). They showed that in this type of test: (a) the multidimensional index l_0 was the sum of the unidimensional indices, and (b) the mean and variance of l_0 were the sum of the unidimensional means and variances, respectively. Next, they used these results to propose a multidimensional standardized index with the form (7.13).

Ferrando (2009) proposed multidimensional extensions of indices (7.16) and (7.17) intended for the FA model. For k common factors, the first index is:

$$M - lco_i = \sum_{j=1}^{n} \left[\frac{X_{ij} - \mu_j - \lambda_{j1}\hat{\theta}_{i1} .... - \lambda_{jk}\hat{\theta}_{ik}}{\sigma_{\varepsilon j}} \right]^2 = -2\ln(\mathbf{X}_i \mid \hat{\theta}_i) + C\,, \tag{7.18}$$

where C is a constant value that does not depend on the trait levels. Under the same assumptions used in (7.16) the expected distribution of (7.18) is χ^2 with $n - k$ degrees of freedom. When $k = 1$, $M - lco$ reduces to the unidimensional index (7.16).

As in the unidimensional case, the second index is a normal approximation to the χ^2 distribution.

$$M - lcz = \sqrt{2M - lco} - \sqrt{2(n-k)-1}\ . \tag{7.19}$$

Little research has been carried out on multidimensional L-B indices and applications are still scarce. Also, whether they make any improvement is not clear. As mentioned earlier, they should be more appropriate for sources of inconsistency that generalize across subtests. And the most likely source of this type is idiosyncratic scale usage (e.g., extreme responding; see Emons, 2009). Sources such as multidimensionality, person unreliability, or certain response biases, however, might well be (at least in part) scale specific. If they are, the multidimensional extensions might be insensitive to overall patterns that show inconsistency on few specific subscales (Schmitt et al., 1999).

In closing, a summary on L-B indices is provided. To start, they have clear limitations and can be improved. Potential lines of improvement are: (a) calibration schemas that take into account the presence of inconsistent patterns, (b) estimation procedures that are more robust and/or use more information, (c) development and use of more accurate reference distributions, possibly obtained via resampling, and (d) development of multidimensional extensions that use more information from the data.

In spite of their limitations, L-B indices are quite useful and the strong criticisms they have received may be largely due to unrealistic expectations. First, both the indices and the models on which they are based are (at best) approximations, and so the indices cannot be expected to closely adhere to a theoretical distribution. Second, inconsistency is a complex phenomenon that has multiple potential sources. So, classification solely based on these indices is a highly error-prone process and it cannot be expected to produce satisfactory results. However, L-B indices are useful as first-step, broad screening tools aimed at flagging potentially problematic patterns. As discussed earlier, once a pattern has been detected, further information must be obtained.

Relative-Variance Indices

L-B indices give an overall idea of the extent to which the observed item scores cannot be well predicted by the IRT model. However, knowing that a pattern is inconsistent does not provide sufficient information about how this inconsistent responding affects the trait level estimate of the respondent. A large negative value of *lz*, for example, is compatible with a small bias in the trait estimate that does not have too much practical relevance, and the opposite can also be true (Meijer & Nering, 1997).

On the basis of a previous proposal by Drasgow, Levine, and McLaughlin (1987), Ferrando (2010) proposed a person fit index for the congeneric model known as *JRV* (jackknife relative variance). This index, which can be used with any of the models considered in this chapter, is based on a deletion approach and uses jackknife estimation (Cook & Weisberg, 1982). Conceptually, the idea is to estimate θ based on different subsets of items, and assess the variance of the resulting estimates. If this variance is large, none of the estimates can be probably trusted. Furthermore, so as to make the index relative, *JRV* is defined as the ratio between the variance of the jackknife trait estimates and the asymptotic variance of the ML estimator (i.e., the model-expected variability; see Ferrando, 2010, for details).

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

$$JRV_i = \frac{Var(\theta_i^*)}{Var(\hat{\theta}_i(ML) \mid \theta_i)}. \tag{7.20}$$

The present proposal is not to use *JRV* as an alternative to the L-B indices but rather as an auxiliary measure that provides complementary information. High values of *JRV* suggest that the estimation of θ is unstable, in the sense that very different estimates might be obtained if certain responses are not considered. Therefore, the point estimate obtained cannot be trusted. On the other hand, small values of *JRV* would indicate that the trait level is consistently estimated by the different item scores.

Expected Behavior of Practical Indices with Different Types of Inconsistency

As a general reference, global indices assess the consistency of the complete pattern, so they are expected to function better with sources of inconsistency that have a global effect on the response vector (Emons, 2009). On the other hand, they are expected to show less sensitivity to sources that affect responses to individual items or small groups of items.

Person Unreliability (Ferrando, 2004; Lumsden, 1977)

Person unreliability can be conceptualized as an individual-differences dimension with two extremes. When going in the direction of high reliability, the response pattern becomes more deterministic or error free, and at the extreme, the pattern behaves according to Guttman's model. In the other direction, the pattern becomes more insensitive to the normative ordering of the items and, at the extreme, the item responses are totally random.

From a person fit point of view, person unreliability produces global observed-expected deviations and is generally well detected by L-B indices (Ferrando, 2004). Highly reliable respondents produce patterns that are too consistent given the stochastic assumptions of the IRT model and that tend to give rise to large and positive values of *lz* and $lz^{(p)}$ (large and negative for *lcz*). On the other hand, the insensitive patterns of unreliable respondents tend to give rise to large and negative values of *lz* and $lz^{(p)}$ (large and positive for *lcz*). The extreme of random responding, however, might be problematic because in this case the items do not provide model-based information for estimating θ so the power is likely to be low. As discussed later, graphical procedures are generally more suitable for detecting this extreme.

Multidimensionality and Idiosyncratic Responding

Waller and Reise (1992) consider that multidimensionality is mostly expected to arise when many items are weakly related to the measured trait. For these items, the influence of the individual's specific factor score outweighs that of his/her common score and this gives rise to idiosyncratic responses. This source of misfit affects specific items or groups of items. So it is difficult to predict whether L-B indices will have enough sensitivity to detect it. If omission of the outlying responses substantially changes the trait estimate, the *JRV* index is expected to be more sensitive.

Faking

Faking can be modeled by assuming a temporal change in the trait level of the individual intended to provide improved scores (Zickar & Drasgow, 1996). The basic point is whether the amount of change tends to remain essentially constant over the different items or whether it tends to vary as a function of the items (i.e., fakeable items vs.

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

faking-resistant items). In the first case, the faked pattern would be elevated but consistent, so faking would not be detected by standard indices. In the second case, provided that the inconsistency was strong enough, faking might be detected (Zickar & Drasgow, 1996).

Recent research (Ferrando & Anguiano-Carrasco, 2013; Zickar & Sliter, 2012) suggests that faking is expected to produce some intra-individual inconsistency in the response pattern. However, the degree of inconsistency is generally subtle and not large enough to be detected with a practical person fit index (Ferrando & Chico, 2001; Reise & Flannery, 1996; Reise & Waller, 2009). As discussed later in this chapter, optimal indices are thought to be more appropriate here.

Acquiescence

Several authors (Curtis, 2004; Reise & Flannery, 1996) have conjectured that acquiescence might be a detectable source of misfit. Ferrando and Lorenzo-Seva (2010) analytically derived some results and arrived at the following predictions. If a fully balanced scale can be obtained, and acquiescence is operating, then (a) the estimated trait level is expected to be essentially correct, but (b) both L-B and *JRV* indices are expected to flag this respondent as inconsistent given the overall large discrepancies between the observed and model-expected item scores. It is much harder to make predictions when a balanced scale is not available because in this case the trait estimate is expected to be biased. If the scale is not balanced at all, acquiescence will probably remain undetected. As discussed later, inconsistency in the case of partially balanced scales is better assessed by using graphical and single-response analyses.

Sabotaging/Malingering

Patterns that can be qualified as sabotaging or malingering have been identified in various data sets (Ferrando, 2012). The trend is that the respondents agree with the most extreme or "difficult" items and disagree with the "easier" items. The degree of inconsistency that this type of responding produces is global and strong and is generally well detected by L-B indices.

Extreme and Middle Responding

Both extreme and middle responding are global sources of misfit. Inconsistency due to extreme responding is expected to be well detected by unspecific indices (Emons, 2009; Ferrando, 2010).

Middle responding produces undifferentiated patterns that make it very difficult to obtain accurate trait estimates. The lack of information results in L-B indices with reduced power for detecting this source and not even the *JRV* is expected to detect instabilities. However, as discussed later, middle responding is easily detected by using graphical procedures.

Specific Indices: Optimal Indices

Levine and Drasgow (1983) proposed a general likelihood ratio test person fit procedure intended to be used to detect specific forms of misfit. The test is based on two likelihoods for a response pattern: (a) the likelihood given a certain model of inconsistent responding, and (b) the likelihood given a model of consistent responding. So the procedure requires a model-based profile of misfit to be specified and, as a result, it is (theoretically) more informative and powerful than the unspecific indices.

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

Consider the likelihoods (7.7), (7.8), and (7.9) for the different models in this chapter. They correspond to the normative model of consistent responding and will be denoted here generically as $L_C\,(\boldsymbol{x}_i \mid \theta)$.

Assume now that an alternative likelihood can be specified for the same pattern based on an IRT modeling of the specific inconsistency to be assessed, and denote this likelihood by $L_{IC}\,(\boldsymbol{x}_i \mid \theta)$. Finally, let $f(\theta)$ be the density of θ. The unconditional likelihoods are:

$$\begin{aligned} P_C(\mathbf{x}_i) &= \int_{-\infty}^{\infty} L_C(\mathbf{x}_i \mid \theta) f(\theta) d\theta \\ P_{IC}(\mathbf{x}_i) &= \int_{-\infty}^{\infty} L_{IC}(\mathbf{x}_i \mid \theta) f(\theta) d\theta, \end{aligned} \tag{7.21}$$

(in practice, the integrals in (7.21) can be approximated with the required precision by using numerical procedures).

The likelihood ratio (LR) is now obtained as:

$$LR = \frac{P_{IC}(\mathbf{x}_i)}{P_C(\mathbf{x}_i)}. \tag{7.22}$$

And the decision rule is to classify $\mathbf{x}_i$ as inconsistent if:

$$P_{IC}(\mathbf{x}_i) \geq \omega\, P_C(\mathbf{x}_i). \tag{7.23}$$

From a purely statistical point of view, test (7.23) is optimal in the Neyman-Pearson sense (hence the name): for a fixed error rate among consistent respondents, no other test has a greater probability of correctly classifying inconsistent respondents. The critical value ω can be interpreted as a cutoff value that controls the ratio between hit rates (proportion of inconsistent respondents classified as such) and false alarm rates (proportion of consistent respondents misclassified as consistent).

The main practical limitation of an optimal index is the specification of the alternative likelihood corresponding to the inconsistency model. For this reason, applications in typical-response measurement are rather scarce and have been limited to the identification of faking. Zickar and Drasgow (1996) modeled the alternative likelihood as a shift of +0.50 to the right of the θ scale for those items that were deemed fakeable while no shift occurred in the remaining items. Ferrando and Anguiano-Carrasco (2013) proposed to obtain both likelihoods in (7.21) by using a partially invariant factor-analytic model that is fitted simultaneously to two data sets: neutral and experimentally induced faking. This second proposal produced better detection rates than many of the procedures reported so far. Thus, further research about its behavior seems warranted.

Like L-B indices, multidimensional versions of the optimal index (7.22) have been proposed (Drasgow, Levine, & McLaughlin, 1991). In Ferrando and Anguiano-Carrasco's (2013) study, the use of the multidimensional extension resulted in a clear improvement in power with respect to the indices based on single scales. However, more research into its behavior is needed.

Graphical Procedures

Most of the procedures proposed so far for the graphical assessment of person misfit (a) are derived from a basic function that can be termed "Person Response Curve" (PRC; Weiss, 1973), and (b) are not intended to be used instead of the scalar-valued indices but rather as useful tools that complement the information provided by these indices (Emons, Sijtsma, & Meijer, 2005; Nering & Meijer, 1998; Sijtsma & Meijer, 2001).

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

The proposal by Weiss (1973) considered the PRC to be the expected score of person *i* as a function of some item difficulty or location scale δ. The two basic assumptions were (a) that the location parameter δ was continuous, and (b) that the PRC was decreasing in δ, so that the more difficult the item was, the lower the expected score of the person on this item.

The general approach proposed in this section is based on the PRC principles discussed earlier and consists of using a nonparametric curve to assess the fit of an expected curve. Both curves are obtained by plotting the item responses of the individual against the ordered item difficulty/location values that, in all the cases, are defined on the θ-continuum.

The graphical representation I propose has three elements: (a) the theoretical or expected person response curve (EPRC), (b) the observed responses, and (c) the empirical or observed person response curve (OPRC). The EPRC is the Weiss curve as defined earlier. The OPRC is the nonparametric smoothed curve that best fits the observed responses, and so does not impose any particular functional form for the curve. Of the several smoothing approaches that can be chosen to fit the OPRC, I propose to use kernel smoothing (KS). KS is widely used, relatively simple, and, when applied at the item level, works well even in comparison with more complex procedures (Härdle, 1990). In accordance with common terminology (Emons, Sijtsma, & Meijer, 2004), both curves will be denoted here generically by $E(S_i|\ \delta)$.

The appropriateness of the individual response pattern is graphically assessed by inspecting the discrepancies between the OPRC and the EPRC. Large discrepancies indicate person misfit. These discrepancies can be general (e.g., curves with opposite trends) or reflect local deviations in certain groups of items (Emons et al., 2004, 2005). Additionally, pointwise confidence intervals can be obtained at the evaluation points and then joined by a line to draw confidence bands on the estimated OPRCs. These confidence bands, which can be obtained via resampling (Emons et al., 2004) or analytically (see Härdle, 1990, section 4.2), provide two important pieces of information: (a) the extent to which the OPRC is well defined across the range of δ considered (which is assessed by the width of the bands), and (b) the regions at which there are significant discrepancies with respect to the EPRCs.

Graphical Procedures for Binary Responses

In the 1PM the EPRC is directly considered to be a function of the item location parameter *b*, so it is defined as:

$$E(S_i \mid b) = \Psi\left(a(\theta_i - b)\right). \tag{7.24}$$

And it is a decreasing one-parameter ogive in which the person trait level θ_i defines the point along the *b* difficulty continuum at which $E(S_i) = 0.5$.

The OPRC can be obtained by using the Nadaraya-Watson KS estimator (see Härdle, 1990):

$$E(S_i \mid b) = \frac{\sum_j^n K(\frac{b - b_j}{h}) X_{ij}}{\sum_j^n K(\frac{b - b_j}{h})}, \tag{7.25}$$

where *K(x)* is the KS function, a nonnegative, continuous, bounded, and (usually) symmetric function that assigns its highest values to points near 0.0 and decreases as it gets further away from 0.0. The parameter *h* is called the bandwidth; it is selected by the user, and controls the amount of smoothing.

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

The definition of the EPRC becomes more complex for the 2PM because in this case the expected item score depends on two parameters, so the b parameter does not order the items identically for each θ.

$$E(S_i \mid b) = \Psi\left(a_j(\theta_i - b)\right). \tag{7.26}$$

As discussed earlier, however, a common finding when the 2PM is fitted to normal-range measures is that the item discriminations do not differ greatly. If they do not, the expected points obtained by (7.26) will not exactly define a line, but they will be tightly clustered around a well-defined decreasing trend. Assuming that this is the case, it is proposed to define the EPRC in the 2PM case as the KS curve that best fits the scatter of expected points obtained by (7.26). So, in this case, the KS estimator (7.25) is used both with the observed points and with the expected points derived from the model predictions.

Graphical Procedures for Continuous Responses

When assessment is based on the congeneric model it is more convenient to work with parameterization (7.6) for two main reasons. First, the EPRC based on (7.6) is decreasing, as it should be. Second, it is more clearly interpretable and the interpretation is equivalent to that of the binary models.

As in the binary case, we shall first consider the τ-equivalent restricted case in which all the discriminations λ are equal. The EPRC in this case can be directly considered to be a function of the location parameter β, and it is defined as:

$$E(S_i \mid \beta) = \left[0.5 + \lambda\theta_i\right] - \lambda\beta. \tag{7.27}$$

Equation (7.27) describes a decreasing straight line in which, as in the 1PM case in (7.24), the person trait level θ_i defines the point along the β continuum at which $E(S_i) = 0.5$ (i.e., the scale midpoint). Note also that the negative slope $-\lambda$ reflects the common discriminating power of the items.

As occurs with the 2PM, in the general congeneric case with different item discriminations:

$$E(S_i \mid \beta) = 0.5 + \lambda_j(\theta_i - \beta), \tag{7.28}$$

the expected values in (7.28) as a function of β will no longer accurately define a line. However, in many applications, the same considerations discussed in the binary case lead us to expect the points to be tightly clustered around a well-defined linear decreasing trend. If this is the case, it is again proposed to define the EPRC as the KS curve that best fits the scatter of expected points.

Graphical Procedures for Graded Responses

In the GRM an item is no longer defined by a single location parameter but by m thresholds. So the graphical representation becomes more complex. In this chapter, attempts are made to obtain a single graphic for each respondent, which leads to the basic initial problem: How can the graded-response items be ordered by some single location value defined on the θ-continuum?

Consider the item-trait regression (7.3), which has been defined as the item response function for a graded-response item. I define the generalized difficulty index (GDI) of

item *j* as the trait value at which the expected score in (7.3) is the response scale midpoint. The GDI is thus defined on the θ-continuum, and has a similar interpretation to the b_j and β_j location parameters in the binary and the congeneric models. Conceptually, GDI*j* is the point on the trait continuum that marks the transition from the tendency to disagree with the item to the tendency to agree with it, and so it can be interpreted as a generalized threshold. Overall, the graphical representation proposed for the GRM case is a single plot for each respondent that displays the item scores as a function of the ordered GDIs.

If the expected scores for an individual are obtained by using his/her trait estimate in Equations (7.2) and (7.3), and then plotted against the corresponding GDIs, the resulting points are not expected to fall onto a single line, not even if all the items have the same discrimination. This is because the spacing and distribution of the thresholds are generally different for different items. So, in the present proposal, the points will only accurately define a curve if, in addition to the restriction of equal discriminations, further restrictions such as those considered by Andrich (1978) are fulfilled (i.e., the thresholds remain invariantly spaced across items, so that they can be shifted left or right but their relative spacing remains the same). Experience suggests that most normal-range typical-response items do not greatly depart from Andrich's (1978) restrictions. For this reason, it is also proposed to define the EPRC in this case as the KS curve (7.25) that best fits the scatter of expected points just described.

These conditions may not be fulfilled in the case of clinical items that measure narrow traits. Reise and Waller (2009) noted that in many applications of this type discriminations varied considerably, the range of thresholds was limited, and threshold values were extreme. They also considered that these results were more likely to be obtained in scales that measure quasi-traits (i.e., traits that are relevant in only one direction). In contrast, the base results that justify our EPRC proposal (similar discriminations, equally spaced thresholds) are more plausible in medium-to-broad bandwidth scales that measure normal-range dimensional traits (as defined in Tellegen, 1988). In any case, the adequacy of the proposal is an empirical question. For the EPRC to be meaningful the scatter must clearly define a decreasing trend and this result must be checked by inspecting the data.

Graphical Assessment of Different Types of Inconsistency

Extreme and Middle Responding

Both extreme and middle responding are better assessed by inspecting the scatter of observed scores. As one of the illustrative examples later in this chapter shows, these sources are generally detected quite easily.

Person Unreliability and Random Responding

Person unreliability mainly affects the amount of dispersion of the observed points around the EPRC. So for over-consistency the observed points are tightly clustered around the EPRC, while for low reliability they are widely scattered. In the binary and graded-response models, person reliability also affects the shape of the OPRC, which becomes steeper for highly reliable respondents and flatter for unreliable respondents. In all the models the extreme of random responding gives rise to a flat OPRC that reflects the total insensitivity of the responses to the normative item ordering (Emons et al., 2005).

Acquiescence

No specific shape for the OPRC can be predicted in this case. However, in balanced and partially balanced scales, inspection of the observed points is expected to show unusually

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

large distances between the two groups of items (positive and negative direction) and the EPRC.

Sabotaging/Malingering

As mentioned earlier, sabotaging/malingering tends to produce an OPRC with a trend that is opposite to that of the EPRC (i.e., increasing instead of decreasing) and that generally can be detected quite easily.

Multidimensionality and Idiosyncratic Responding

Because it is a specific source of misfit, this type of responding is better assessed by inspecting the distances of the individual points from the EPRC. When misfit affects small groups of items (e.g., lower-level facets), the OPRC is likely to show local deviations with respect to the EPRC that can be detected with the help of the confidence bands.

Faking

Faking is unlikely to be detected by graphical procedures. Both the OPRC and the EPRC are expected to be elevated with respect to the generally unknown "true" curves. However, they are expected to be essentially consistent with each other. These predictions are illustrated in one of the empirical examples later in this chapter.

Scalar Valued Item-Level Indices

Assessment of misfit at the level of single item response will be discussed by using a basic distinction from regression diagnostics: outliers and influential observations (Cook & Weisberg, 1982; Zijlstra, van der Ark, & Sijtsma, 2007). In the present context, outliers are item responses that are highly unlikely given the model parameters and the estimated trait level of the respondent. Influential observations, on the other hand, are item responses that have a disproportionate influence on the estimated trait level of the respondent. More operatively, an item response can be considered to be influential when the estimated trait level of the respondent changes substantially when this response is deleted from the data (Ferrando, 2010; Zijlstra, van der Ark, & Sijtsma, 2007). In general, outliers need not be influential observations (Cook & Weisberg, 1982; Meijer & Nering, 1997).

Outliers

The most immediate and well-known indices of this type are scaled residuals between the observed and the model-expected item response. And the most common way of scaling the residual is to divide the raw observed-expected difference by the expected conditional standard deviation given the person estimate (e.g., Smith, 1990). The scaled residual then takes a *z*-score form, and its values are interpreted with reference to the standard normal distribution (Karabatsos, 2000).

For the one- and two-parameter models, the best-known scaled residual is the individual standardized residual proposed initially by Wright (1977) for the Rasch model:

$$z_{ij(BINARY)} = \frac{X_{ij} - P_j(\hat{\theta}_i)}{\sqrt{P_j(\hat{\theta}_i)(1 - P_j(\hat{\theta}_i))}}. \tag{7.29}$$

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

For the GRM case, the corresponding index is that proposed by Wright and Masters (1982) for the Rating Scale Model:

$$z_{ij(GRADED)} = \frac{X_{ij} - E(X_{ij} \mid \hat{\theta}_i)}{\sqrt{Var(X_{ij} \mid \hat{\theta}_i)}}, \tag{7.30}$$

where the second term in the numerator is the expected score (7.3) evaluated with the person estimates, and the conditional variance is given by:

$$Var(X_j \mid \hat{\theta}_i) = \left[\sum_r r^2 P_{jr}(\hat{\theta}_i)\right] - \left[E(X_j \mid \hat{\theta}_i)\right]^2. \tag{7.31}$$

Finally, the linear counterpart of the residuals (7.29) and (7.30) is the statistic proposed by Bollen and Arminger (1991) and by Ferrando (2010):

$$zc\ e_{ij} = \left(\frac{X_{ij} - \mu_j - \lambda_j \hat{\theta}_i(ML)}{\sqrt{Var(e_{ij})}}\right) = \frac{e_{ij}}{\sqrt{Var(e_{ij})}}, \tag{7.32}$$

where:

$$Var(e_{ij}) = \sigma^2_{\varepsilon j} - \frac{\lambda_j^2}{\sum_j^n \frac{\lambda_j^2}{\sigma^2_{\varepsilon j}}}. \tag{7.33}$$

From a graphical point of view, indices (7.29), (7.30), and (7.32) are standardized distances between the observed item score and the EPRC. Therefore, they are expected to be particularly useful for those types of inconsistency that are assessed on the basis of inspection of the observed response points: acquiescence, multidimensionality, and idiosyncratic responding.

Indices of types (7.29) and (7.30) have been criticized for two main reasons (Karabatsos, 2000; Smith, 1990). The first is the same as that of the L-B indices: the same response pattern is used to estimate the person parameters and analyze the fit. The second concerns the chosen reference distribution. The indices are transformed discrete variables that cannot be well approximated by a continuous distribution such as the standard normal.

Index (7.32) explicitly takes into account that the ML trait estimate is used instead of the "true" trait level. Furthermore, the item response is assumed to be continuous. It then follows that, if the congeneric model is correct and the item parameters are known, index (7.32) should follow a standard normal distribution. As discussed earlier, however, the congeneric model is only an approximation, and so is the reference distribution.

While the theoretical limitations so far discussed are relevant, item-level indices are generally used in practice to trace unexpected item responses in patterns that have been detected as potentially inconsistent by a global index so that insight can be gained into the causes of the misfit. They are not intended to be used as strict inferential measures, and the reference distribution is only used as a useful reference. As Smith (1990) discussed, for these purposes, the indices are possibly useful enough as they are.

Influential Observations

The developments that are now discussed are general and apply to all the IRT models considered here. To start, according to the definition of influential observations, the most direct measure of the influence of item *j*'s score is:

$$D_i(j) = \theta_i(ML) - \theta_i^{(-j)}(ML), \tag{7.34}$$

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

that is, the change in the ML trait estimate when item *j*'s score is deleted from the data. While (7.34) is indeed direct, it is difficult to interpret because of the lack of reference values. This limitation can be improved by adopting a resampling procedure proposed by Zijlstra, van der Ark, and Sijtsma (2007). In our case it consists of (a) randomly deleting a single item score and obtaining the change estimate using (7.34), and (b) repeating the process 1,000 times and establishing a confidence interval for the change values. If the estimated change corresponding to item *j* lies outside the boundaries of the confidence interval, this score can be regarded as influential.

A second way of making (7.34) more interpretable is to transform it with an appropriate scaling. Ferrando (2010) proposed a pseudo-standardized scaled measure that he termed $Dz_i(j)$, and that is based on the properties of the ML estimator:

$$Dz_i(j) = \frac{\hat{\theta}_i(ML) - \theta_i^{(-j)}(ML)}{\sqrt{Var(\hat{\theta}_i(ML) \mid \theta_i)}}. \tag{7.35}$$

Conceptually, *Dz* is a modification of Cook's distance (Cook & Weisberg, 1982), which informs of the change that takes place in the estimated trait level with respect to the variability that is expected in the estimation based on this particular response pattern.

Researchers must know how to proceed when influential observations have been detected. One possibility is not to trust the trait estimate or to retest the respondent (Meijer & Sijtsma, 2001). A second possibility is to re-estimate the trait level. One option within this second approach (Meijer & Nering, 1997) is simply to eliminate the influential scores and estimate θ on the basis of the remaining item scores. A more elaborate option is to use a robust procedure such as the biweight that downsizes the impact of the influential points. This second approach should not be automatically applied. It is expected to work well and provide a more valid estimate when only a few scores are identified as both outlying and influential and in which a convincing rationale for the discrepancies can be found. In other cases the modified trait estimate will probably continue to be completely meaningless.

Application

The examples in this section are applications in the personality domain intended to illustrate some of the points discussed in this chapter. In all cases, and as recommended, overall model-data fit and item fit were assessed before person fit. However, given the illustrative purposes of the examples we shall only focus on person fit results.

Example 1: A comparison of GRM-based and congeneric-based person fit results

The CTAC (Spanish acronym for Tarragona Questionnaire of Anxiety for Blind People; Pallero et al., 1998) is intended for blind and visually impaired people. It contains 35 items with a five-point response format and attempts to measure anxiety related to visual loss in a range of everyday situations. The CTAC items are generally non-extreme and have moderate discriminating power. So, as expected, both the GRM and the congeneric model fit the data about equally well. Given this result, the first example will illustrate the comparability of the person fit results based on both models.

We shall first discuss some overall results. First, the product-moment correlation between both sets of ML trait estimates was $r = 0.99$. Second, the correlation between $lz^{(p)}$ and *lcz* was $r = -0.70$ (negative as expected). However, the agreement between both indices was considerably higher than the correlation indicates. The dispersion of points was considerably lower at the extreme of inconsistency where the respondents that are flagged

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

as inconsistent by both indices are concentrated. So the most inconsistent respondents would be flagged with one index or the other.

The first individual illustration corresponds to respondent number 393, who was flagged as potentially inconsistent according to the L-B indices. The $lz^{(p)}$ estimate obtained under the GRM calibration was $lz^{(p)} = -3.01$ whereas the *lcz* based on the linear model was 1.70. In both cases, however, the relative variance measures suggested that the variability between the trait estimates was not excessive. The *JRV* estimates were 1.36 (GRM) and 1.38 (congeneric model).

Figure 7.1 shows the graphical assessment based on the GRM (panel a) and the congeneric model (panel b). The similarity between both graphics is remarkable and they both clearly lead to the same diagnosis. The EPRC is decreasing, as it should be. However, the OPRC is virtually flat, which suggests that the respondent is largely insensitive to the normative ordering of the items. This type of result might be caused by a high degree of person unreliability, inapplicability of the trait (i.e., low traitedness) or, in the extreme, random responding. Further analyses do not seem necessary in this case, and it seems reasonable to assume that the trait estimate corresponding to this respondent cannot be validly interpreted.

The second illustration corresponds to respondent number 628. The L-B estimates in this case were: $lz^{(p)} = -3.90$ (GRM) and $lcz = 3.59$ (congeneric model). The corresponding relative variance indices were $JRV = 2.84$ (GRM) and $JRV = 2.04$ (congeneric model). There is close agreement between the results obtained from both models, which suggests that: (a) the response pattern of number 628 is highly inconsistent and (b) the trait estimates are rather unstable.

The two panels in Figure 7.2 again show the graphical assessment based on both models, and, as in the previous illustration, they agree very closely. In spite of the large global estimates, it is clear that the OPRC and the EPRC do not greatly differ in this case (compare with the previous assessment). However, the inspection of the graphs reveals interesting results. Note first that individual number 628 is probably an extreme responder, as most of the responses are 1 or 5 (0 and 1 in the transformed congeneric scale). Second, note that there are three potential outliers at the top right of the graph. Item-level analysis clearly identified these points with standardized residual values of 3.07, 2.69, and 2.64 (GRM-based *z*), and 2.56, 2.43, and 2.38 (congeneric-model-based *zc*). They are marked with an asterisk in Figure 7.2.

Contrary to what would be expected given the relatively high *JRV* values, the three outliers were not influential observations (*Dz* estimates of around 0.30 for the three points). Overall, the pseudo-standardized influential estimates were moderate for most of the items and none of them exceeded 1.65. Given the results it is not clear how to proceed with the trait estimate of this respondent. The trait estimate based on the complete pattern was the same in both models (about 0.65) and the jackknife estimates ranged from 0.60 to 0.75. So, in spite of the large amount of potential inconsistency detected by the global indices, the trait estimate in this case might well be essentially valid and interpretable.

Example 2: Experimentally induced faking

The second example shows how difficult it is to detect faking in real data sets when unspecific person fit procedures are used. The measure in this case is the Psychoticism (P) scale of the Eysenck Personality Questionnaire revised (EPQ-R; Eysenck, Eysenck, & Barrett, 1985), which consists of 32 binary items. The EPQ-R was administered on two occasions. At Time 1 the participants were asked to respond under the standard instructions provided in the manuals. At Time 2 they received faking-inducing instructions (try to give a good impression so that they will be given a job they really want).

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

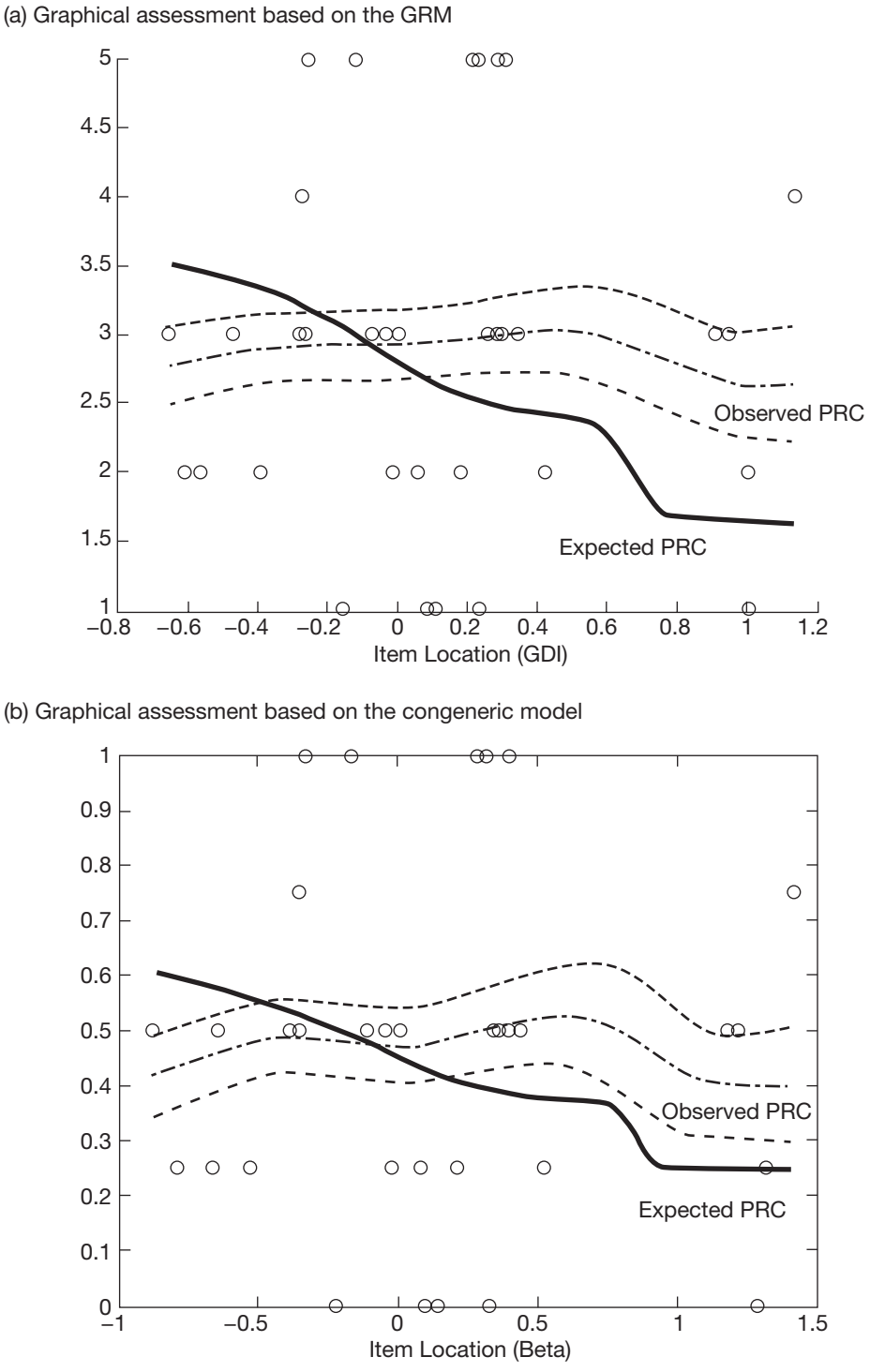


Figure 7.1 Graphical analysis of respondent number 393 based on the GRM (upper panel) and the congeneric model (lower panel). Example 1.

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

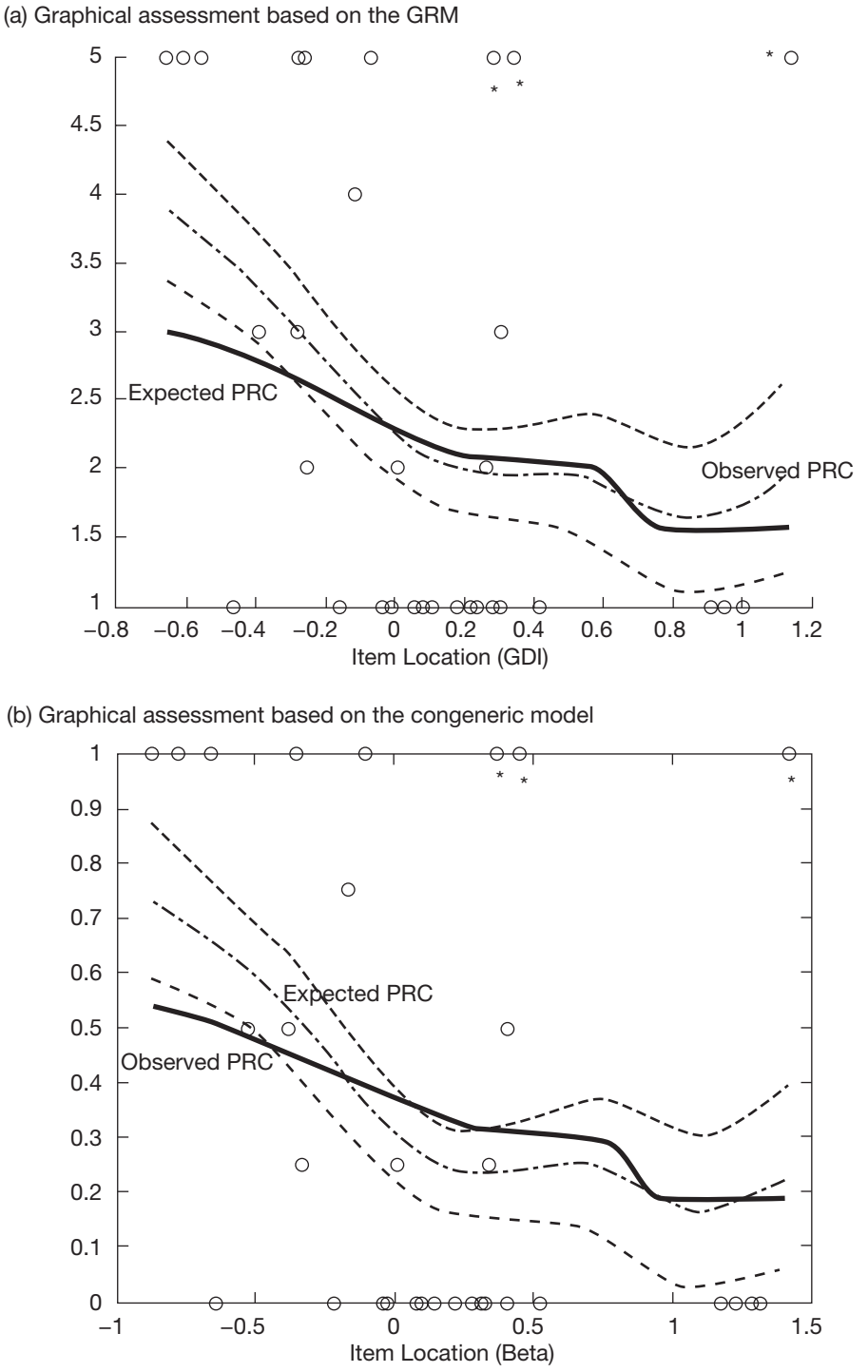


Figure 7.2 Graphical analysis of respondent number 628 based on the GRM (upper panel) and the congeneric model (lower panel). Example 1.

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

In the present illustration, the P items were calibrated using the data obtained in the neutral administration. The 2PM fit the data quite well, and so the estimated parameters were taken as fixed and known. Then on both occasions ML trait estimates were obtained from these item parameters.

For illustrative purposes, the response patterns given by respondent number 156 under both conditions are now assessed. First, a dramatic change in the ML trait estimate is observed. The estimate obtained in neutral conditions is $\hat{\theta}_{156} = 1.30$, which reflects a rather high level of P. However, the estimate obtained under faking good conditions is $\hat{\theta}_{156} = -3.68$, a very large decrease that goes in the expected direction: toward a lower P level (which is far more socially desirable). In spite of this decrease, however, the *lz* statistic was unable to detect the faking behavior of this respondent. The *lz* values were 1.57 (neutral condition) and 0.13 (faking condition), which would lead us to conclude that this participant responded quite consistently on both occasions.

The graphical assessment of these response patterns is shown in Figure 7.3 and helps to explain the result. The considerable difference in elevation between both sets of curves reflects the change in the estimated trait level. However, (a) the profile of both curves is similar, and (b) the OPRC and the EPRC generally agree in both cases. Overall, the results agree with the view that faking behavior generally produces a rather consistent elevation (a decrease in this case) of the scores that is unlikely to be detected by using practical indices. However, when the optimal person fit procedure discussed earlier was applied, the value of the *LR* statistic (7.22) was 1.52, a value that suggests that this respondent is a potential faker.

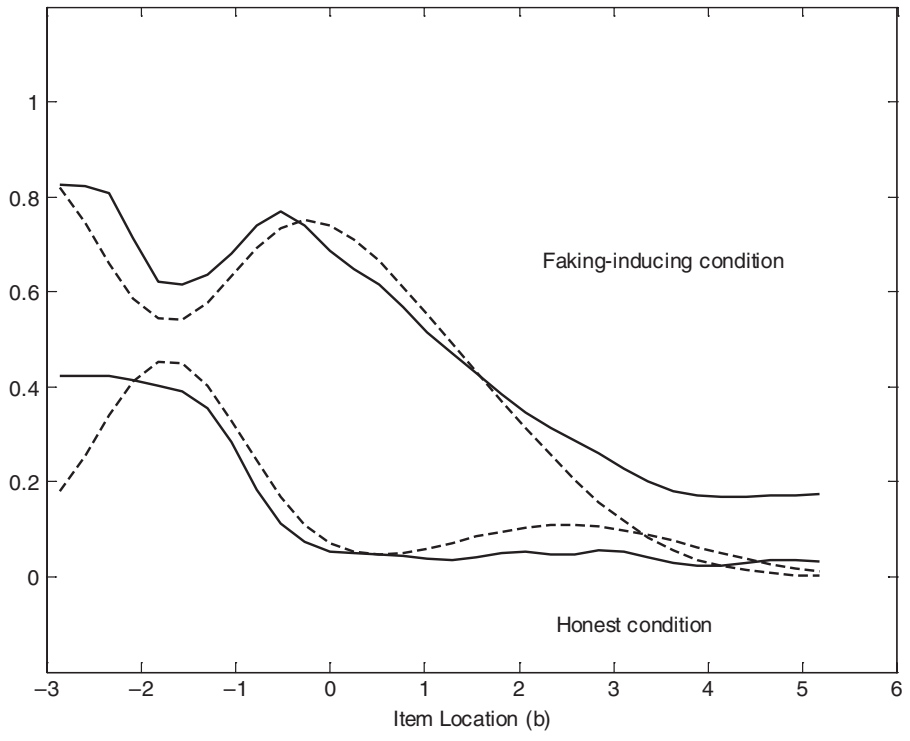


Figure 7.3 Graphical analysis of respondent number 156 based on the 2PM under faking-inducing and honest responding conditions. Example 2.

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

Example 3: Acquiescence on a partially balanced scale

The last example shows how the combined use of practical indices, graphical procedures, and item-level indices can detect acquiescence when the responses are based on a partially balanced scale. The measure in this case was a 35-item extraversion scale in which 25 items measured in the direction of the extraversion pole and the remaining 10 measured in the direction of the introversion pole. All the items were positively worded and used a five-point Likert format. The sample size was 480, and the data was well fitted by the GRM.

Response pattern 201 had a $lz^{(p)}$ estimate of –3.76, and the *JRV* estimate was 1.66. The first index suggests that this respondent answered inconsistently. However, the second suggests that the variability between the trait estimates is not much higher than the expected variability of the estimate based on the full response pattern. So no highly influential item scores are expected in this case.

As discussed earlier, acquiescence is thought to be better assessed when single-item discrepancies are used and the scatter of points is inspected. Inspection of the standardized residuals in Equation (7.30) showed large outliers (seven of which had an absolute value higher than 1.65) that closely corresponded to the 10 reverse items. Furthermore, all these residuals were negative. This is to be expected in the case of acquiescence: when the raw score in the introversion-worded items is reversed, the resulting score is lower than expected given the trait estimate of this respondent. As was also expected, inspection of the *Dz* influential indices in (7.35) showed that all the values corresponding to the reverse items were positive (when the lowered item score is omitted, the trait estimate based on

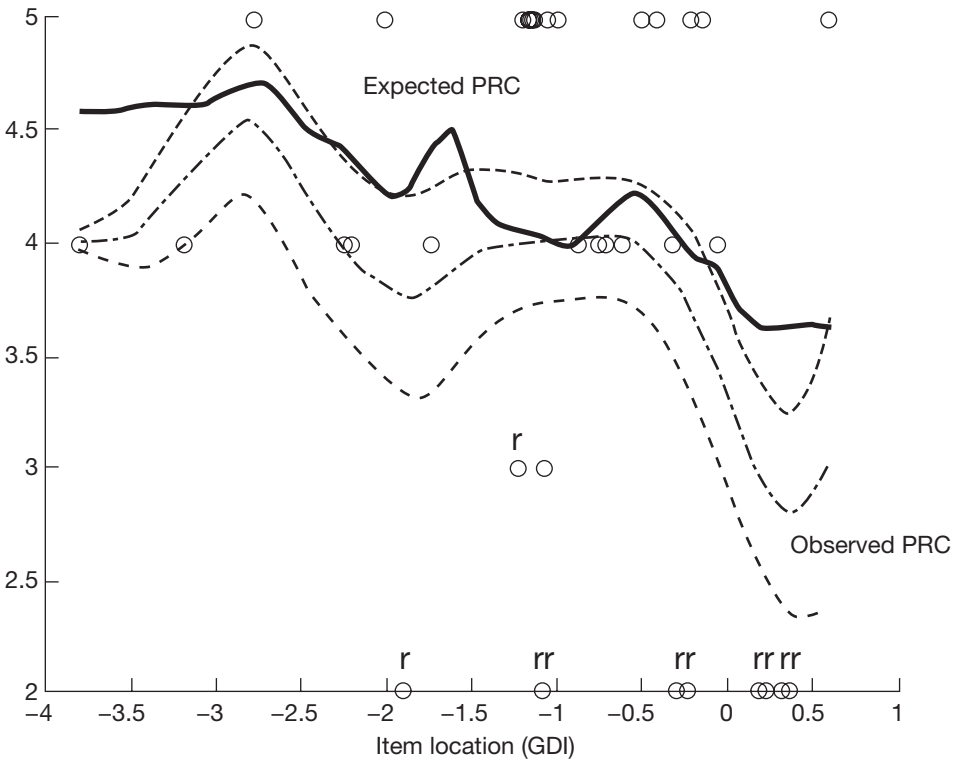


Figure 7.4 Graphical analysis of respondent number 201 based on the GRM. Example 3.

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

the remaining items increases). However, the *Dz* values were rather low, and none of them was higher than 0.5.

Figure 7.4 shows the graphical assessment of response pattern number 201 and, to help interpret the results, the items that were reversed are marked with an "r." The OPRC locally deviates from the OPRC at the right end of the graph, possibly because of the cluster of reversed items at the bottom right. Overall, however, what is most clearly seen in the graph are the large negative outliers, which correspond to the items that were reverse-scored.

It is hard to decide the extent to which the trait estimate of this respondent can be trusted. If the scale were fully balanced it would be assumed that the effects on the positive and the reverse set would cancel each other out, so the estimate based on the full pattern would have been reasonably correct (see Ferrando, 2010). However, the scale is only partially balanced, so it is likely that the trait estimate based on the full pattern is upwardly biased.

Future Directions

I begin the discussion with a caveat. This chapter has tried to provide a comprehensive approach to person fit in typical-response measures. However, in spite of its general purpose, this chapter reflects the views of the author, and part of the text is devoted to procedures the author himself has developed. So the chapter cannot be properly considered to be a review of existing person fit procedures. Person fit is a wide domain of research containing multiple approaches that reflect different views. So the interested reader is encouraged to explore alternative procedures and perspectives.

In recent decades typical-response psychometric applications have become more rigorous and, at present, most of them are model based. Therefore, an acceptable model-data fit is a basic requirement that has to be met before the test can be used to score individuals. However, in most applications at present, assessment of fit finishes once the test scores have been obtained: all the scores are then assumed to be valid indicators of the trait levels and, therefore, they are interpreted and/or used for selection purposes or in validity studies. In this chapter I have tried to make it clear that this assumption is not warranted. So my position is that person fit should always be assessed before the scores are interpreted or used. I have also made it clear that person fit procedures have considerable limitations and need to be improved in the future. Even so, they are useful, and the potential they have to improve measurement justifies their use in any typical-response application.

Experience suggests that recommendations such as this are only widely used if well-developed user-friendly software is readily available. So, in closing, I shall provide some discussion on noncommercial programs that implement the procedures discussed here.

WPerfit (Ferrando & Lorenzo-Seva, 2000) is a Windows program that computes L-B global indices based on the 1PM and the 2PM. It also implements graphical procedures, and obtains the PRCs corresponding to these models. L-B indices for the binary models can also be obtained with the R (http://cran.r-project.org) packages IRTOYS and MIRT. MIRT is able to compute both the unidimensional and multidimensional versions of the *lz* index.

For the GRM case, L-B indices can be obtained with the R program PERSONz (Choi, 2010), and again with MIRT. PERSONz allows cutoff values to be obtained via simulation. MIRT computes both the unidimensional and multidimensional versions of $lz^{(p)}$. Finally, the unidimensional and multidimensional indices proposed by Ferrando for the congeneric model can be obtained with the program FACTOR (Lorenzo-Seva & Ferrando, 2013).

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

The other indices and procedures discussed in this chapter are implemented in ad hoc programs that, at best, would be useful for methodologically oriented researchers. So a great deal of work is needed if the present proposal is to be widely used in applied research. And perhaps the best future line of action is to develop a comprehensive, user-friendly program that computes global and item-level indices together with clear and powerful graphical displays. In the meantime, the cited software allows applied researchers in the personality and attitude domains to undertake the most basic forms of person fit assessment. If this chapter convinces them to incorporate this assessment (albeit at the most basic level) in their applications, it will have fulfilled an important aim.

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2*, 449–460.

Armstrong, R. D., Stoumbos, Z. G., Kung, M. T., & Shi, M. (2007). On the performance of lz statistic in person fit measurement. *Practical Assessment, Research & Evaluation, 12*.

Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. In P. V. Marsden (Ed.), *Sociological methodology 1991* (pp. 235–262). New York: Basil Blackwell.

Chang, H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response function in polytomously scored item response models. *Psychometrika, 59*, 391–404.

Choi, S. W. (2010). PERSONz: Person misfit detection using the lz statistic and Monte Carlo simulations. *Applied Psychological Measurement, 34*, 457–458.

Clark, J. M. (2010). *Aberrant response patterns as a multidimensional phenomenon: Using factor-analytic model comparison to detect cheating*. Doctoral dissertation, University of Kansas.

Conrad, K. J., Bezruczko, N., Chan, Y., Riley, B., Diamond, G., & Dennis, M. L. (2010). Screening for atypical suicide risk with person fit statistics among people presenting to alcohol and other drug treatment. *Drug and Alcohol Dependence, 106*, 92–100.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. London: Chapman & Hall.

Culpepper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Applied Psychological Measurement, 37*, 201–225.

Curtis, D. D. (2004). Person misfit in attitude surveys: Influences, impacts and implications. *International Education Journal, 5*, 2125–2144.

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*, 159–177.

Dodeen, H., & Darabi, M. (2009). Person-fit: Relationship with four personality tests in mathematics. *Research Papers in Education, 24*, 115–126.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59–79.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171–191.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

Egberink, I. J. L., & Meijer, R. R. (2011). An item response theory analysis of Harter's Self-Perception Profile for Children or why strong clinical scales should be distrusted. *Assessment, 18*, 201–212.

Emons, W. H. M. (2009). Detection and diagnosis of person misfit from patterns of summed polytomous item scores. *Applied Psychological Measurement*, *33*, 599–619.

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypothesis about the person-response-function in person-fit analysis. *Multivariate Behavioral Research, 39*, 1–35.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local and graphical person-fit analysis using person-response functions. *Psychological Methods, 10*, 101–119.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*, 105–120.

Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. T. (1985). A revised version of the Psychoticism scale. *Personality and Individual Differences, 6*, 21–29.

Ferrando, P. J. (2002). Theoretical and empirical comparisons between two models for continuous item responses. *Multivariate Behavioral Research, 37*, 521–542.

Ferrando, P. J. (2004). Person reliability in personality measurement: An item response theory analysis. *Applied Psychological Measurement, 28*, 126–140.

Ferrando, P. J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research, 42*, 481–508.

Ferrando, P. J. (2009). Difficulty, discrimination and information indices in the linear factor-analytic model for continuous responses. *Applied Psychological Measurement, 33*, 9–24.

Ferrando, P. J. (2010). Some statistics for assessing person-fit based on continuous-response models. *Applied Psychological Measurement, 34*, 219–237.

Ferrando, P. J. (2012). Assessing inconsistent responding in E and N measures: An application of person-fit analysis in personality. *Personality and Individual Differences*, *52*, 718–722.

Ferrando, P. J., & Anguiano-Carrasco, C. (2013). A structural model-based optimal person fit procedure for identifying faking. *Educational and Psychological Measurement, 73*, 173–190.

Ferrando, P. J., & Chico, E. (2001). Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection scales. *Educational and Psychological Measurement, 61*, 997–1012.

Ferrando, P. J., & Lorenzo-Seva, U. (2000). WPerfit: A program for computing parametric person-fit statistics and plotting person response curves. *Educational and Psychological Measurement*, *60*, 479–487. (Available at http://psico.fcep.urv.es/utilitats/wperfit/)

Ferrando, P. J., & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 427–448.

Härdle, W. (1990). *Applied nonparametric regression*. London: Chapman & Hall.

Hofstee, W. K. B., Ten Berge, J. M. F., & Hendricks, A. A. J. (1998). How to score questionnaires. *Personality and Individual Differences, 25*, 897–910.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory. Application to psychological measurement*. Homewood: Dow Jones-Irvin.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109–133.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement, 1*, 152–176.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277–298.

Levine, M. V., & Drasgow, F. (1983). Appropriateness measurement: Validating studies and variable ability models. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 109–131). New York: Academic Press.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics, 4*, 269–290.

Levy, P. (1973). On the relation between test theory and psychology. In P. Kline (Ed.), *New approaches in psychological measurement* (pp. 1–42). New York: Wiley.

Li, M. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21*, 215–231.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lorenzo-Seva, U., & Ferrando, P. J. (2013). FACTOR 9.2: A comprehensive program for fitting exploratory and semi-confirmatory factor analysis and IRT models. *Applied Psychological Measurement*, *37*, 497–498. (Available at http://psico.fcep.urv.es/utilitats/factor/)

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement, 1*, 477–482.
Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's *lz** index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics, 37*, 57–81.
McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.
Meijer, R.R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*, 3–8.
Meijer, R.R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8*, 72–87.
Meijer, R.R., Egberink, I.J.K., Emons, W.H.M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's Self-Perception Profile for Children. *Journal of Personality Assessment*, *90*, 1–14.
Meijer, R.R., & Nering, M.L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement*, *21*, 321–336.
Meijer, R.R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107–135.
Mellenbergh, G.J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29*(3), 223–237.
Molenaar, I.W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*, 75–106.
Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*, 171–189.
Nering, M.L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*, 121–129.
Nering, M.L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*, 321–336.
Nering, M.L., & Meijer, R.R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement, 22*, 53–69.
Olsson, U. (1979). On the robustness of factor analysis against crude classification of observations. *Multivariate Behavioral Research, 14*, 485–500.
Pallero, R., Ferrando, P.J., & Lorenzo-Seva, U. (1998, July). *Questionnaire Tarragona of anxiety for blind people*. Paper presented at the IX International Mobility Conference, Atlanta.
Reise, S.P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*, 213–229.
Reise, S.P., & Due, A.M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*, *15*, 217–226.
Reise, S.P., & Flannery, W.P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education*, *9*, 9–26.
Reise, S.P., & Waller, N.G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, *65*, 143–151.
Reise, S.P., & Waller, N.G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48.
Reise, S.P., Waller, N.G., & Comrey, A.L. (2000). Factor analysis and scale revision. *Psychological Assessment*, *12*, 287–297.
Reise, S.P., & Widaman, K.F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods, 4*, 3–21.
Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. (Psychometrika Monograph No. 17). Iowa City: Psychometric Society.
Schmitt, N., Chan, D., Sacco, J.M., McFarland, L.A., & Jennings, D. (1999). Correlates of person-fit and effect of person-fit on test validity. *Applied Psychological Measurement, 23*, 41–53.

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

Sijtsma, K., & Meijer, R.R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66*, 191–207.

Smith, R.M. (1990).Theory and practice of fit. *Rasch Measurement Transactions, 3*, 78–80.

Snijders, T.A.B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*, 331–342.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality, 56*, 622–663.

van Krimpen-Stoop, E.M.L.A., & Meijer, R.R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement, 26*, 164–180.

Waller, N.G., & Reise, S.P. (1992). Genetic and environmental influences on item response pattern scalability. *Behavior Genetics, 22*, 135–152.

Weiss, D.J. (1973). *The stratified adaptive computerized ability test*. Research report 73–3. Minneapolis: University of Minnesota.

Woods, C., Oltmanns, T.F., & Turkheimer, E. (2008). Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychological Assessment*, *20*, 159–168.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97–116.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis*. Chicago, IL: Mesa Press.

Zickar, M.J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71–87.

Zickar, M.J., & Sliter, K.A. (2012). Searching for unicorns: Item response theory-based solutions to the faking problem. In M. Ziegler, C. MacCann, & R.D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 113–130). New York: Oxford University Press.

Zijlstra, W.P., van der Ark, L.A., & Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behavioral Research*, *42*, 531–555.

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7

This page intentionally left blank

Downloaded By: 10.3.97.143 At: 21:11 06 Dec 2023; For: 9781315736013, chapter7, 10.4324/9781315736013.ch7