The Use of the l(z) and l(z)* Person-Fit Statistics and Problems Derived From Model Misspecification

Meijer, Rob R.; Tendeiro, Jorge N.

Published in:
Journal of Educational and Behavioral Statistics

DOI:
10.3102/1076998612466144

Journal of Educational and Behavioral Statistics
2012, Vol. 37, No. 6, pp. 758–766
DOI: 10.3102/1076998612466144

The Use of the l_z and l_z* Person-Fit Statistics and Problems Derived From Model Misspecification

Rob R. Meijer
Jorge N. Tendeiro
Department of Psychometrics and Statistics, University of Groningen

We extend a recent didactic by Magis, Raîche, and Béland on the use of the l_z *and* l_z* *person-fit statistics. We discuss a number of possibly confusing details and show that it is important to first investigate item response theory model fit before assessing person fit. Furthermore, it is argued that appropriate distributions are only a first step for practical use of person-fit statistics.*

Keywords: *item response theory, person fit, model fit, aberrant response patterns, validity scales*

> As all detective stories remind us, many of the circumstances surrounding a crime are accidental and misleading. Equally, many of the indications to be discerned in bodies of data are accidental and misleading. To accept all appearances as conclusive would be destructively foolish, either in crime detection or data analysis. To fail to collect all appearances because some—or even most—are only accidents would, however, be gross misfeasance deserving (and often receiving) appropriate punishment.
>
> (Tukey, 1977, p. 3)

In a recent paper in *Journal of Educational and Behavioral Statistics*, Magis, Raîche, and Béland (2012) provided a didactic on the use of the two person-fit statistics l_z and l_z*. Both statistics are standardized likelihood-based statistics that are used to determine the likelihood of an item score pattern under a specified item response theory (IRT; Embretson & Reise, 2000) model. We applaud the use of person-fit statistics in educational and psychological measurement and we certainly think that a didactic like the one provided by Magis et al. (2012) can further popularize the use of these statistics. However, we also think that in their study *there were some places where more clarification could help readers* when applying person-fit statistics in practice. The aim of the present study is to add a number of empirical results for these statistics and to provide some further suggestions that may guide the further direction of person-fit research.

The main message in Magis et al. (2012) was to *allow researchers to reproduce the rather complicated calculations for* l_z* and to show that l_z* should be preferred over l_z (Drasgow, Levine, & Williams, 1985) because its distribution has a better approximation to the assumed standard normal distribution than l_z. Although it was originally suggested that l_z followed a standard normal distribution, several studies showed that this does not hold when an estimated trait value is used (e.g., van Krimpen-Stoop & Meijer, 2001). Snijders (2001), therefore, proposed an adapted version of l_z denoted l_z* that corrects for the estimated trait value. Because Snijders's (2001) work was rather technical, in Magis et al. (2012) a useful, more practical, guide to calculate l_z^* was provided. Furthermore, an empirical example was given, in which the distributions of l_z and l_z* were compared with the normal distribution using empirical data of a language assessment test and the 2-parameter logistic model (2PLM) to describe these data. It is here that some confusion may arise for practitioners that may want to apply these person-fit statistics. Magis et al. (2012) showed empirical distributions where the left tail of the distribution of the l_z statistic better fits the standard normal curve than the l_z* statistic for larger negative values than -2, and l_z* better fits the standard normal for values between -2.0 and 1.0. It is the left tail of the distribution that is most relevant because here are the values that reflect misfit of an item score pattern with an IRT model. As Magis et al. (2012) state, the worse fit of the l_z* statistic for the extreme values may be due to a general misfit of the IRT model to these language assessment data. This is plausible because for the cognitive language assessment data the 3-parameter logistic model (3PLM) seemed to be more appropriate than the 2PLM. In short, this empirical example is not very illustrative with respect to the main point Magis et al. (2012)—correctly—would like to discuss: Use l_z* instead of l_z. We further clarify how the distributions of these statistics are behaving using real empirical data, we show the effect of model misfit, and we further discuss a strategy to apply person-fit statistics in practice.

There are two points we would like to stress before we report our empirical results. (1) In general, we think that it is a good strategy to first investigate whether the data can be described in a reasonable way by a particular IRT model. That is, a researcher should first investigate model and item fit. In Magis et al. (2012), the 2PLM is chosen to estimate the likelihood functions, but there is no check on whether the 2PLM gives a reasonable description of the data (at least, it is not reported). As a result, the misfit of an item score pattern may be due to the general misfit of the IRT model to the data, or due to unexpected answering behavior of a particular person. Because model misfit also affects the distribution of a person-fit statistic as was illustrated in Magis et al. (2012), this is an important point in practical research projects. (2) There is, in our view, an essential difference between person fit and item fit. It is fairly easy to remove an item from a test because it does not fit the model, for example, because of violations of local independence. However, withholding a test score from a test taker because the generated score pattern is unlikely under an IRT model is often problematic. In many test situations, the assessor should then have additional

information about the underlying reasons of misfit. This does not imply that sampling distributions are not important, but that a good story is perhaps more important when applying these statistics in practice. We will elaborate on this below.

Examples

We analyzed two scales of the Dutch Personality Questionnaire–Junior (Dutch: Junior Nederlandse Persoonlijkheidsvragenlijst, NPV-J; Luteijn, van Dijk, & Barelds, 2005). The total scale consists of 105 mostly positively formulated items and is intended to determine how adolescents between 9 and 15 years of age judge their own behavior on five scales. We selected the Inadequacy (IN, 28 items) and the Social Inadequacy (SI, 13 items) scales because these scales had the best psychometric properties (Weekers & Meijer, 2008). Scoring was originally done on a 3-point scale (*Agree, ?, Disagree*) but because the instructions of the NPV-J discourage the use of the *?* response, and because we were afraid that many adolescents would choose the *?* category, we used a 2-point scale (*Agree*, *Disagree*). The answer *Agree* was scored as 1 and the answer *Disagree* was scored as 0. Data were collected from 866 persons who attended primary and secondary education in the east of the Netherlands.

Data were collected for research purposes, that is, data were only collected for a project that was aimed at investigating the usefulness of different types of IRT models to fit personality scores. Thus, the participants had no stakes at filling out the questionnaires and motivation may have been low for some students. Note that this is a context that is not uncommon in test practice. For example, when data are collected to assess the quality of a test, or when a test is administered to assess the quality of a group and not the individual, motivation problems may be present. In fact, when administering the questionnaires it was clear that some students did not make a very serious attempt to fill out the questionnaire. Thus, we expected that in this data set there were patterns that represent random response behavior or other types of idiosyncratic behavior, for example, choosing the yes answer for each item irrespective of the content of the item.

Results

From a psychological and data analytical perspective, it is interesting to first study the distribution of the total scores. Both distributions were positively skewed: For the IN scale the mean score was 6.23 ($SD = 5.19$), and for the SI scale the mean score was 5.01 ($SD = 3.17$).

Before we show the person-fit statistics distributions, we first report the fit of the 2PLM. Trait estimates and item parameters were estimated using the program IRTPRO (Cai, Thissen, & du Toit, 2011). The 2PLM was chosen because it seems to be a natural choice to describe the answers to a personality questionnaire. In Table 1, the item mean scores (p values), and the a-parameters and b-parameters are given for the IN scale from the 2PLM together with the S-X^2

TABLE 1
*IN Scale: Mean Item Scores (*p *values), Item parameters, and S-X^2 Probabilities*

Item	Mean	*a* Parameter	*b* Parameter	Prob. S-X^2
In1	.13	1.16	1.97	.69
In4	.24	0.81	1.63	.88
In6	.20	1.03	1.64	.95
In8	.11	2.50	1.47	.74
In13	.15	0.67	2.88	.81
In14	.14	2.30	1.35	.75
In19	.17	0.85	2.12	.66
In28	.06	3.49	1.68	.92
In32	.26	0.90	1.39	.14
In34	.33	1.00	0.86	.76
In36	.19	1.50	1.31	.90
In38	.09	1.61	1.94	.51
In48	.32	1.12	0.86	.99
In50	.30	1.25	0.90	.61
In52	.40	0.86	0.53	.73
In54	.08	2.85	1.65	.64
In57	.28	1.42	0.92	.95
In59	.17	1.29	1.55	.66
In66	.09	2.85	1.58	.92
In70	.33	1.78	0.61	.66
In72	.19	1.13	1.57	.93
In75	.15	2.18	1.30	.67
In91	.11	1.27	2.05	.96
In93	.19	3.11	1.02	.93
In96	.24	1.72	1.00	.77
In98	.56	0.80	–0.33	.31
In100	.50	0.95	0.00	.74
In102	.24	2.17	0.91	.48

item level diagnostics (Orlando & Thissen, 2003); for the SI scale, this information is presented in Table 2. For both scales, for almost all items these diagnostics point at good fit. Furthermore, we inspected the marginal fit and the standardized local dependence X^2 statistic obtained from IRTPRO (see also, Chen & Thissen, 1997). This X^2 statistic is computed by comparing the observed and expected frequencies in each of the two-way cross tabulations between responses to each item and each of the other items. For the IN scale, of the 406 item pairs, there were only 9 X^2 statistics pairs larger than 5 and 1 larger than 10. The IRTPRO manual suggests that values larger than 5 are in a gray area of fit; values larger than 10 point at a violation of local dependence. For the SI scale, similar results were obtained.

TABLE 2
*SI scale: Mean Item Scores (*p *values), Item Parameters, and S-*X^2 *probabilities*

Item	Mean	*a* Parameter	*b* Parameter	Prob. S-X^2
Si21	.68	1.21	–0.80	.37
Si22	.32	1.04	0.91	.21
Si23	.44	2.32	0.22	.10
Si25	.27	1.38	0.97	.06
Si26	.25	2.03	0.91	.02
Si44	.65	0.44	–1.49	.38
Si51	.24	1.38	1.10	.34
Si62	.45	2.26	0.16	.03
Si79	.19	1.76	1.24	.17
Si80	.35	0.51	1.30	.30
Si85	.50	2.21	0.01	.74
Si89	.53	1.56	–0.08	.31
Si105	.14	1.50	1.65	.79

Some further inspection of the characteristics of the IN dataset revealed that the item *b* parameters were clustered within a limited range (with one exception the *b* parameters were all positive). This results in suboptimal conditions to classify item scores as aberrant or normal. Meijer, Molenaar, and Sijtsma (1994) showed that the larger the spread between the item difficulties, the higher the detection rate. This can easily be understood by realizing that almost all person-fit statistics weigh the item scores with the item difficulty. When there is not much spread in the item difficulties, the person-fit values for different item score patterns will be similar. In the hypothetical case that all item difficulties are the same, every configuration of item scores will have the same likelihood.

The good fit of the 2PLM results in an accurate empirical distribution of the l_z*. In Figure 1A we depicted the person-fit distributions for the IN scale and in Figure 1B we depicted the distribution for the SI items. Most interesting are the differences in the left tail of the distributions. In Figure 1B, it can be seen that the left tail of the l_z distribution is too light, whereas the left tail of the l_z* distribution is very close to the standard normal. For example, for a 5% significance level taking $l_z^* < -1.65$, the l_z* distribution almost perfectly follows the normal distribution. The light tail of the l_z distribution results in a conservative classification of misfitting patterns. Note that these results are a good illustration of the superiority of the l_z* distribution in contrast to the results presented by Magis et al. (2012). In Figure 1B, the distributions for the SI subtest are given and similar conclusions hold as for the IN subtest, although the tail of the l_z* distribution is somewhat lighter than the tail of the normal distribution. Although the SI scale

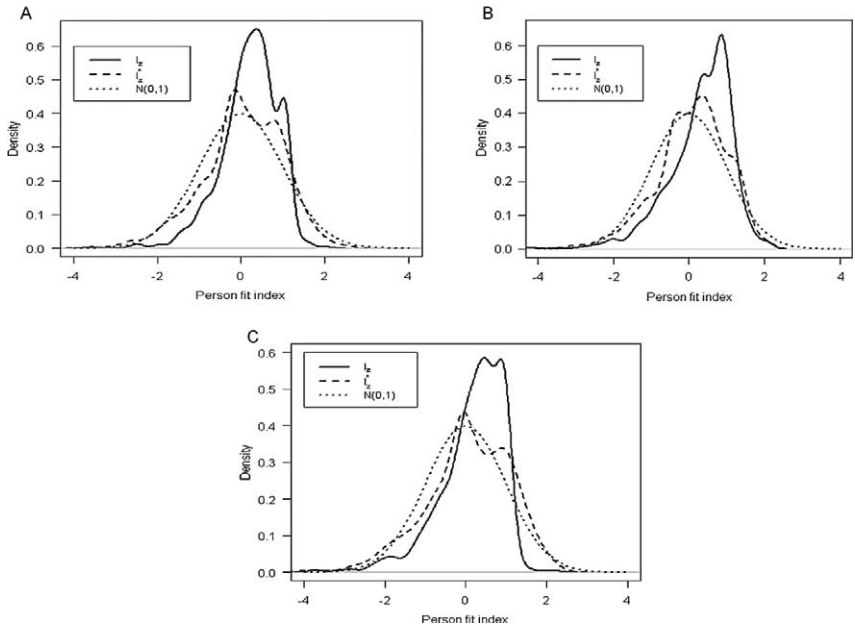

FIGURE 1. *A. Distributions of person-fit statistics for the IN subtest (28 items) under 2PLM. B. Distributions of person-fit statistics for the SI subtest (13 items) under 2PLM. C. Distributions of person-fit statistics for the IN items under 1PLM.*

only consists of 13 items, the left-tail distribution of l_z* is remarkably close to the normal. To see what happens when we have a very short test, we calculated the distributions for two scales (Scholastic Competence, SC, and the Physical Appearance scale, PA) consisting of 6 items each measuring self-concept (see Meijer, Egberink, Emons, & Sijtsma, 2008). Note that Snijders's (2001) results were based on asymptotic theory, that is, based on long tests. As can be seen in Figures 2A and 2B, for short tests, the distributions are far from the normal density. As an alternative for short tests, one may use a complete enumeration procedure, that is, one may calculate the probability for each pattern and obtain an exact distribution (Molenaar & Hoijtink, 1990; Tendeiro & Meijer, 2012a). However, it is debatable whether there is enough reliable information in a short test that warrants the use of inspecting the configuration of item scores.

To study what happens when we fit the wrong IRT model, we restricted all *a*-parameters to a constant. Using IRTPRO, all *a*-parameters were set to 1.29. Inspecting the item fit of both the IN items and the SI items under this model resulted in several misfitting items. The effect on the distribution of the person-fit statistics is shown in Figure 1C. It is interesting to observe that the left tail of both the l_z and the l_z* distributions are thicker than when using the 2PLM that better fits the data,

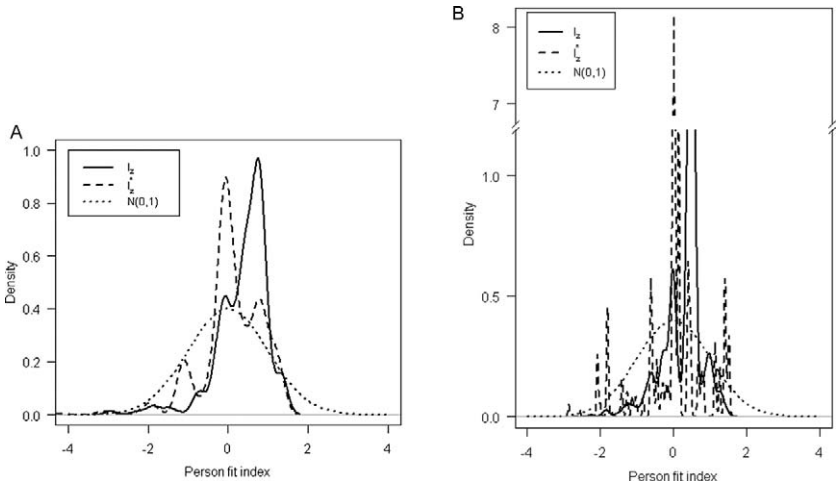

FIGURE 2. *A. Distributions of person-fit statistics for the SC scale of the SPPC (6 items). B. Distribution of person-fit statistics for the PA scale SPPC (6 items).*

although for the l_z* distribution the differences are not large. The heavier tails reflect more misfitting response patterns, due to model misfit. Inspecting this figure, it may be argued that in the left tail the distribution of the l_z statistic provides a better approximation to the normal, but this is because the distribution is inflated due to many misfitting response patterns due to model misfit. In fact, this figure resembles the figure in the Magis et al. (2012) paper and illustrates why we should be careful in classifying item score patterns as aberrant based on the wrong model.

Interpretation of Misfit: The Biggest Challenge

Thus far, we only discussed statistical issues, but the most important issue is that there is little evidence that person-fit scores mean anything psychologically or are useful for invalidating scores. In the present study, there was not a large percentage of persons who filled out the IN or SI scale randomly: The percentage of extreme negative l_z* scores was as expected under the 2PLM. Similar results were found recently by Ferrando (2011) . One of the few studies that tried to validate person-fit scores was Meijer et al. (2008). They used interviews with both children and teachers to clarify the meaning of poor person fit. This kind of research is of critical importance in demonstrating the potential value and validity of person fit. Another interesting example was given in Conrad et al. (2010) who identified patients with atypical item score patterns on a depression questionnaire. These patterns were characterized by high scores on severe indicators of depression such as high scores on questions about suicidal thoughts and low scores on less severe indicators of depression like sleeping problems. Of course,

we are not against using better statistical tools, but we think that using l_z* instead of l_z or similar statistics discussed in Meijer and Sijtsma (2001; see also Klauer, 1995 and Tendeiro & Meijer, 2012b) is only a first step to the more fruitful research area of person-fit. We need more studies that show how these statistics can be used to improve educational and psychological measurement. To paraphrase the motto of this article, "we need to collect more appearances."

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Cai, L., Thissen, D., & du Toit, S. (2011). *IRTPRO 2.1 for Windows*. Chicago, IL: Scientific Software International.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs: Using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

Conrad, K. J., Bezruczko, N., Chan, Y.-F., 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.

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.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Ferrando, P. J. (2011). Assessing inconsistent responding in E and N measures: An application of person-fit analysis in personality. *Personality and Individual Differences*, *52*, 718–722.

Klauer, K. C. (1995). The assessment of person fit. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 97–110). New York, NY: Springer Verlag.

Luteijn, F., van Dijk, H., & Barelds, D.P.H. (2005). *NPV-J: Junior Nederlandse Persoonlijkheidsvragenlijst. Herziene handleiding 2005* [NPV-J: Dutch Personality Questionnaire–Junior: Professional manual (revised)]. Amsterdam, The Netherlands: Harcourt Assessments.

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's *l(z)** index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, *37*, 57–81.

Meijer, R. R., Egberink, I. J. L., 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*, 227–238.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, *18*, 111–120.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indexes. *Psychometrika*, *55*, 75–106.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289–298.

Snijders, T. A. B. (2001). Asymptotic null distribution of person-fit statistics with estimated person parameter. *Psychometrika*, *66*, 331–334.

Tendeiro, J. N., & Meijer, R. R. (2012a). A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement*, *36*, 420–442.

Tendeiro, J. N., & Meijer, R. R. (2012b). The probability of exceedance as a nonparametric person-fit statistic for tests of moderate length. *Manuscript submitted for publication*.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of a person-fit statistic in fixed- and computerized adaptive testing. *Applied Psychological Measurement*, *23*, 327–345.

Weekers, A. M., & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models: An illustration with a Dutch dominance and unfolding personality inventory. *European Journal of Psychological Assessment*, *24*, 65–77.

Authors

ROB R. MEIJER is professor at the University of Groningen, Faculty of Social and Behavioral Sciences, Department of Psychometrics and Statistics, Grote Kruisstraat 2/1, 9712 TS, Groningen, the Netherlands; e-mail: r.r.meijer@rug.nl. His research interests are in test theory and item response theory and their applications in educational and psychological measurement.

JORGE N. TENDEIRO is assistant professor at the University of Groningen, Faculty of Social and Behavioral Sciences, Department of Psychometrics and Statistics, Grote Kruisstraat 2/1, 9712 TS, Groningen, the Netherlands, e-mail: j.n.tendeiro@rug.nl. His research interests are in three-way data analyses and in item response theory.

Manuscript received May 11, 2012
Revision received August 15, 2012
Accepted September 4, 2012