Cluster bias: Testing measurement invariance in multilevel data

Jak, S.

Publication date
2013

CHAPTER 1

Measurement bias and multidimensionality; an illustration of bias detection in multidimensional measurement models

Abstract Restricted factor analysis can be used to investigate measurement bias. A prerequisite for the detection of measurement bias through factor analysis is the correct specification of the measurement model. We applied restricted factor analysis to two subtests of a Dutch cognitive ability test. These two examples serve to illustrate the relationship between multidimensionality and measurement bias. We conclude that measurement bias implies multidimensionality, whereas multidimensionality shows up as measurement bias only if multidimensionality is not properly accounted for in the measurement model.

Based on: Jak, S., Oort, F.J. & Dolan, C.V. (2010). Measurement bias and multidimensionality; an illustration of bias detection in multidimensional measurement models. *Advances in Statistical Analysis, 94*, 129-137.

INTRODUCTION

In the presence of measurement invariance, systematic differences between observed test scores are attributable to true differences in the trait(s) that the test measures. A test is measurement invariant with respect to V, if the following conditional independence holds:

$$f_1(X \mid T = t, V = v) = f_2(X \mid T = t), \quad (1)$$

where X is a set of observed variables, T is a set of attributes measured by X, and V is a set of variables other than T, possibly violating measurement invariance. Function f_1 is the conditional distribution function of X given values of t and v, and f_2 is the conditional distribution function of X given t. If the conditional independence does not hold (i.e., if $f_1 \neq f_2$), the measurement of T by X is said to be biased with respect to V. In the presence of measurement bias, differences between observed test scores may not represent true differences between respondents.

The principle of conditional independence (PCI) was introduced by Mellenbergh (1989) to define item bias (or differential item functioning), with X representing a test item, T a latent trait, and V some group membership. Yet Mellenbergh emphasized the generality of the definition: X, T, and V may be measured on the nominal, ordinal, interval or ratio level, they may be latent or manifest, and their relationships may be linear or nonlinear. In their review of statistical methods for the detection of measurement bias, Millsap and Everson (1993) distinguished between latent variable methods (with latent T) and observed variable methods (with observed T), but they only considered group membership as possible V. Oort (1991) showed that a whole range of measurement issues can be subsumed under the PCI. Relevant measurement issues only differ in what is substituted for X (e.g., item responses, test scores), T (e.g., one or more latent traits), and V (e.g., other items, other latent traits, group membership, time of measurement occasion, socio-demographic variables). Oort called variables V potential violators of unbiased measurement (hence the symbol V). Meredith (1993) used the PCI to define weak measurement invariance and factorial invariance across populations defined by V, and called V a selection variable.

Structural equation modeling (SEM) with latent variables provides flexible means to test measurement invariance, i.e., measurement issues related to the PCI-based definition of unbiased measurement can be investigated using SEM. Most typically, the X variables are observed variables (item scores or test scores) and the T variables are continuous latent variables. The V variables can be group membership in multigroup data, time index in longitudinal data (see King-Kallimanis, Oort & Garst (2010) for an example), or any other

variable, observed or latent. Different SEM methods to detect measurement bias with respect to each of these types of V have been proposed.

If measurement bias is investigated with respect to a nominal V representing groups (e.g., treatment versus control group, men versus women), then we can use multigroup factor analysis (MGFA) with structured means (Sörbom, 1974). In the multigroup method, specific manifestations of bias can be investigated by testing across group constraints on intercepts (uniform bias) and factor loadings (nonuniform bias); see Vandenberg and Lance (2000) for a review. Similarly, measurement bias in longitudinal data (e.g., response shift) can be investigated using longitudinal factor analysis (Oort, 2005).

Another way to detect bias, with respect to any variable (e.g., age, gender, personality trait, attitude, mood), is by conducting restricted factor analysis (RFA) as proposed by Oort (1992, 1998). In the RFA method, uniform bias can be investigated by testing the significance of direct effects of exogenous variables (V) on the observed variables (X). In effect, the RFA method is equivalent with using multiple indicator multiple cause (MIMIC) models to detect measurement bias (Muthén, 1989), the only difference being that in MIMIC models the V variables have causal effects on the T variables, whereas in the RFA method V and T variables are merely associated. Advantages of RFA (and MIMIC analysis) over multigroup factor analysis (MGFA) are that it is not necessary to categorize continuous V variables into groups, and that bias can be investigated with respect to several violators simultaneously.

A prerequisite for the detection of measurement bias through any of these SEM methods is the correct specification of the measurement model. The definition of unbiasedness based on PCI features distributions of X conditional on T. This requires the relationship between X and T, including the dimensionality of T, to be correctly specified. Misspecification of the dimensionality of T in the measurement model may lead spurious bias results (Ackerman, 1992).

In this paper, we present two examples of measurement bias detection through RFA. We focus on the specification of the measurement model, and discuss explicitly the relationship between multidimensionality and measurement bias.

METHOD

The RFA method is used to study measurement invariance of the "Q1000 Capaciteiten Hoog" with respect to age and gender. This is a commercial test, designed to measure cognitive abilities of highly educated people (Meurs HRM, Woerden, The Netherlands). The test consists of seven subtests, with a total of 137 dichotomous items (scored 0 for incorrect, 1 for correct). The test was administered to 1617 respondents (961 men and 656

women, 17 to 63 years of age, m = 37.9, sd = 9.0) as part of a selection procedure for a traineeship in Dutch government. All respondents were highly educated (BA level at least). Here we present the results for two subtests, Mathematical ability and Spatial visualization ability. Prior to investigating measurement bias, we first established the measurement model. Subsequently, we applied the RFA method to investigate bias with respect to gender and age.

ESTABLISHING THE MEASUREMENT MODEL

We first fitted a one-factor model in both subtests. Standardized residuals and modification indices (MIs, this is equivalent to using Lagrange Multiplier tests; Muthén & Muthén, 2006) were used to guide specification search. To guard against capitalizing on chance, the MIs were tested at a Bonferroni adjusted level of significance (nominal alpha of 5% was divided by $p(p-1)/2$, where p is the number of items in the subtest). We only permitted modifications that were amendable to substantive interpretation.

DETECTING MEASUREMENT BIAS

Once we established the measurement models, we added gender and age to the model as exogenous variables. Gender and age were allowed to correlate with each other and with the ability factor(s), but all direct effects of gender and age on the test items were fixed to zero. Measurement bias was evaluated by testing these zero direct effects, using MIs. If the largest of the MIs was significant at a Bonferroni adjusted alpha level (nominal alpha of 5% was divided by pq, where p and q are numbers of items and exogenous variables), the direct effect was set free to be estimated. The associated item was then considered biased. This procedure was repeated until none of the remaining fixed direct effects was significant (at a re-adjusted level of significance, i.e., dividing nominal alpha by $pq - r$, where r is the number of direct effects set free).

STATISTICAL ANALYSIS

As the items of the ability tests are dichotomous, we fitted our models to a matrix of tetrachoric correlations, using weighted least squares with adjusted mean and variance (WLSMV) as implemented in Mplus 4.2 (Muthén & Muthén, 2006). WLSMV provides asymptocically correct standard errors and an adjusted χ^2 statistic (Muthén, du Toit and Spisic 1997). All MIs and χ^2 difference tests were re-scaled to improve the approximation of the χ^2 distribution (Satorra & Bentler, 2001).

In addition to the adjusted χ^2 statistic, the root mean squared error of approximation (RMSEA) and the expected cross validation index (ECVI) were used as measures of

overall goodness-of-fit (Browne & Cudeck, 1993). RMSEA values smaller than 0.05 indicate close fit, and values smaller than 0.08 are still considered satisfactory. Confidence intervals around the RMSEA values and ECVI values were calculated with the freely available computer program NIESEM (Dudgeon, 2003).

RESULTS

MATHEMATICAL ABILITY

Mathematical ability is measured with 12 worded, four-choice math problems. Although the overall goodness-of-fit of the one-factor model was reasonable ($\chi^2 = 329.55$, $df = 48$, $p < .01$, RMSEA= .060 [90% CI: .053, .067], ECVI = .241 [90% CI: .208, .279]), significant MIs identified correlated residuals. All items with correlated residuals were at the end of the test. Apparently, time constraints caused respondents to hurry through the last part of the test, so that the results were affected by speed as well as mathematical ability. We added a second factor, labeled "Speed", to account for the extra shared variance in the last six items. The fit of this two-factor model is close ($\chi^2 = 63.50$, $df = 43$, $p = .02$, RMSEA= .017 [90% CI: .007, .026], ECVI = .083 [90% CI: .072, .099]).

Using this measurement model, we added gender and age as exogenous variables (Figure 2). We found a positive correlation between gender and mathematical ability ($r = 0.34$), indicating higher mathematical ability for men, and a negative correlation between age and speed ($r = -0.20$), indicating that older people are slower, which may have affected their test performance. Two items showed bias. Age had a significant direct effect on Item 1 ($\beta = .12$), indicating that the item is easier for older people: In a subgroup of equally able respondents, older respondents perform better on Item 1. Item 2 was found to be biased with respect to both age ($\beta = -.12$) and gender ($\beta = -.13$): For respondents with equal ability, this item was easier for women, and easier for younger people.

We did not find an immediate explanation for Item 1, which was about chicken farmers and their relative numbers of chickens. Item 2 was a worded problem about employees' preferences of what to do at an upcoming office party. To solve the item, one must assume that half of the male employees prefer dancing over bowling. Perhaps the older male respondents have been distracted more than other respondents by the unusual gender role behaviour.

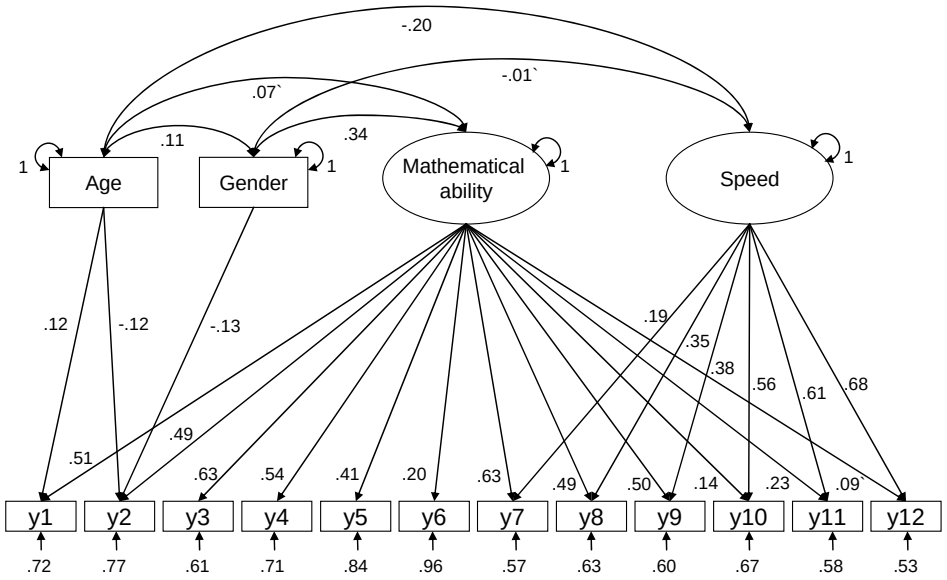

Figure 1 Mathematical ability measured by worded problems.

Notes: All figures denote standardized parameter estimates; apostrophes indicate non-significance; N = 1617; model fit: $\chi^2 = 103.79$, df = 58, $p < .01$, RMSEA= .022 [90% CI: .015, .029], ECVI = .122 [90% CI: .108, .143].

SPATIAL VISUALIZATION ABILITY

The Spatial visualization ability test consists of 17 items. Each item pictures a three-dimensional cube with different patterns on each of its planes. Through mental rotation, respondents have to choose from four options which other cube is a rotation of the first cube.

The overall goodness-of-fit of the one-factor model is reasonable: $\chi^2 = 750.64$, $df = 95$, $p < .01$, RMSEA= .065 [90% CI: .061, .070], ECVI = .537 [90% CI: .485, .594]). However, MIs identified 15 covariances among the item residuals of three subsets of items. Inspection of item content showed that the three groups of items differed in the number of mental rotations needed to solve the items. We modeled this property by adding three factors to the general ability factor, hypothesizing that different mental capacities are required to solve problems that require different numbers of rotations. The fit of this four-factor model was good: $\chi^2 = 133.02$, $df = 87$, $p < .01$, RMSEA= .018 [90% CI: .012, .024], ECVI = .165 [90% CI: .148, .187]).

We added gender and age as exogenous variables to the revised measurement model (Figure 2). Significant positive correlations between gender and general visual-spatial ability (r = .15), specific single rotation ability (r = .12), and double rotation ability (r = .13) indicated that men do slightly better than women. Negative correlations between age and general visual-spatial ability r = -.24), single rotation ability (r = -.18) and triple rotation ability *(*r = -.10) seemed to indicate that the associated skills deteriorate with increasing age. None of the items was found to be biased with respect to age or gender.

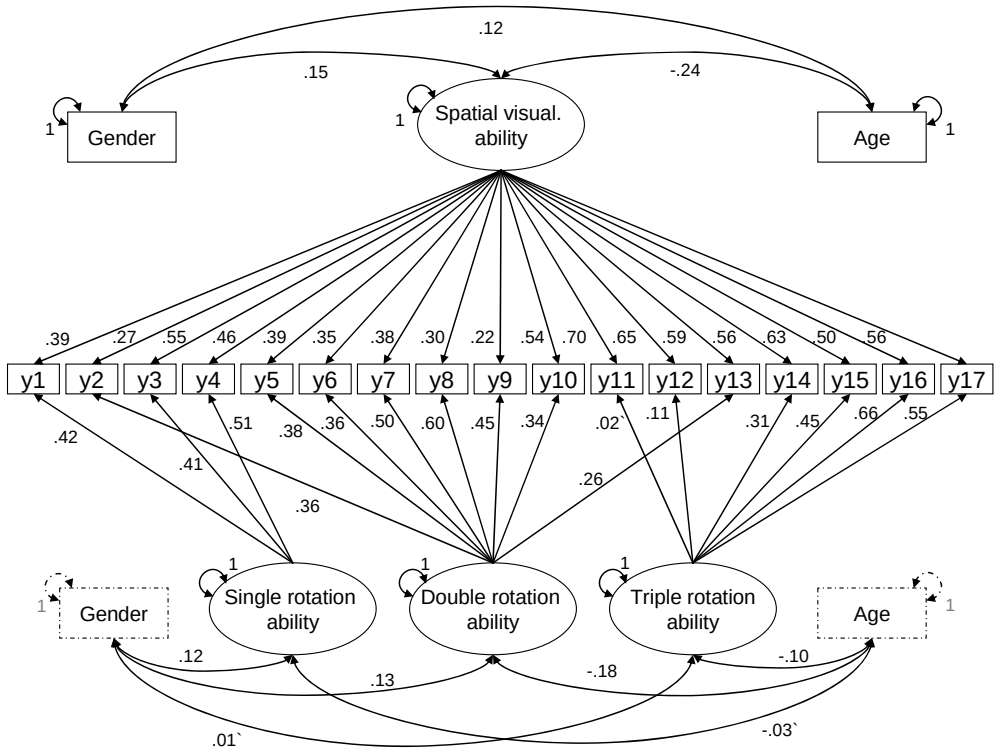

Figure 2 Spatial visualization ability measured by cube rotation problems.

Notes: All figures denote standardized parameter estimates; apostrophes indicate non-significance; for visual clarity, residual variances are not shown, and variables gender and age are pictured twice; N = 1617; model fit: χ^2 = 165.54 with df = 107, p < .05, RMSEA= .018 [90% CI: .013, .023], ECVI = .206 [90% CI: .187, .231].

DISCUSSION

We applied RFA to detect measurement bias with respect to age and gender to two subtests of a Dutch cognitive ability test. We also applied the MGFA method to the cognitive ability data, categorizing age into two age groups and conducting separate

analyses to investigate bias with respect to gender and age. Here, the MGFA and RFA methods yielded very similar results, but the MGFA method does have some disadvantages. In our example, gender and age were correlated (men were older). When we use MGFA to separately investigate bias with respect to gender and age then it might be difficult to distinguish gender bias from age bias. Investigation of gender and age group bias simultaneously in MGFA would involve the comparison of at least four smaller groups (younger women, older women, younger men, older men). Besides complicating the procedure and the interpretation of the results, this also means less precise parameter estimates and loss of statistical power.

Limitations of the RFA method generally come from the measurement bias definition being far more general. For example, in the RFA method T is operationalized as a continuous latent variable, whereas in the definition T can be a discrete latent variable, as in latent class analysis (also incorporated in SEM; Muthén & Muthén, 2006), or T can be an observed variable, as in some of the older bias detection methods such as the Mantel-Haenszel procedure (Holland & Thayer, 1988) and the logistic regression procedure (Swaminathan & Rogers, 1990). Furthermore, in the RFA method only linear conditional independence can be tested, and the method is not readily suited to detect nonuniform bias (although the RFA method can be extended with latent moderated structures; see Barendse, Oort & Garst, 2010). In the MGFA method nonuniform bias can be investigated by testing across group constraints on factor loadings. Still, when we applied the MGFA method to our cognitive ability data we did not find any nonuniform bias.

In the present research we relied on modification indices for model modification, and we tested these at a Bonferroni adjusted level of significance to prevent chance results. Saris, Satorra, & Van der Veld, (2009) suggested to use modification indices in combination with the expected parameter change, and to take the statistical power of the modification index into account as well. This is generally worthwhile, but does not lead to other results in our examples, as the model modifications were already justified substantively and we checked whether the modifications changed the parameter estimates substantially.

In practice it may be difficult to find the true cause of apparent bias, because there may be many possible violators of the measurement model operating simultaneously. Even if all possible violators are known, it will not be possible to operationalize and measure all possible causes of measurement bias. For example, in the worded math problem about office parties we conjecture that the apparent sex and age bias is really caused by the unusual gender role behaviour in the text of the worded problem. As we have no measure of "familiarity with unusual gender role behaviour" available, we can only detect bias with respect to sex and age. Researchers of measurement bias should be aware of this problem, and always try to investigate bias with respect to as many possible violator variables as

available. One of the advantages of the RFA method is that bias can be detected with respect to multiple possible violators simultaneously.

MEASUREMENT BIAS AND MULTIDIMENSIONALITY

The present examples serve to illustrate the relationship between measurement bias and multidimensionality. In both examples we rejected the one-dimensional factor model in favour of a multidimensional factor model. In the first example, if we ignored the speed factor, we found age bias in the last items of the test, which would have been difficult to interpret. In the multidimensional model it is clear that the last items (also) measure speed and that age is correlated with speed. In the second example, the specific rotation factors that vary in their correlations with gender and age could have been mistaken for bias in the associated items. In one of the other Q1000 cognitive ability tests, a 37-item vocabulary test, measurement bias detection yielded multiple items that favoured younger respondents (results not shown here). Inspection of item content showed that these biased items all inquired after the meaning of words with English origin. The biasing factor was therefore taken to be familiarity with English language, which is assumed to be inversely related with age.

In general, the interpretation of apparent measurement bias involves reflection on possible biasing factors. In the one-dimensional model, all items are really affected by two factors: the single common factor and an item-specific residual factor, as in Spearman's (1928) original "two-factor theory". If all residual variance was really only random error variance then measurement bias would be absent by definition. But if the residual variance also contains structural variance then this may stem from a biasing factor. If multiple items in a test are affected by the same biasing factors, these factors may surface as additional common factors, as was the case with speed in the mathematical ability test, the specific rotation factors in the spatial-visual test, and English language familiarity in the vocabulary test. However, if the residual factors do not share any structural variance, then the hypothesis of unidimensionality will not be rejected, although measurement bias may still be present. Oort (1991) used the definition of measurement bias to define unidimensionality as the absence of measurement bias with respect to any variable that might be relevant in whatever context the test is used. Following Lord and Novick's (1968) notion of "complete latent space", we can define k-dimensionality as the number of dimensions of T that is needed to achieve statistical independence of all items X. Modeling all k dimensions guarantees the absence of measurement bias.

With the RFA method, if we operationalize the biasing factor as one of the variables V, we can detect bias with respect to the nuisance factor itself. In the mathematical ability example, we might consider speed to be a biasing factor, and the effects of the speed factor on Items 7 through 12 as measurement bias. Instead of the speed factor as an

additional *T* in a multidimensional measurement model, the speed factor then features as a latent *V* in a model with a unidimensional *T*. This once more shows that multidimensionality and measurement bias really address the same problem. Measurement bias in a unidimensional model may disappear in a multidimensional model. The other way around, misspecification of the dimensionality of *T* in the measurement model may lead to spurious findings of bias.

In conclusion, measurement bias and multidimensionality are related, but not equivalent. Measurement bias implies multidimensionality, but multidimensionality shows up as measurement bias only if multidimensionality is not properly accounted for in the measurement model.