pp. 68-78 - *Bulletin de Méthodologie Sociologique* (59 rue Pouchet, F 75017 Paris), March 1998, N. 58

ON GOING RESEARCH RECHERCHE EN COURS

DETECTING ABERRANT RESPONSE PATTERNS IN MULTI-ITEM SCALES: SOME NONPARAMETRIC INDICES AND A COMPUTER PROGRAM

by

Edith D. de Leeuw & Joop J. Hox
(Department of Education, University of Amsterdam,
Wibautstraat 4, NL-1091 GM Amsterdam, Netherlands;
email hox@educ.uva.nl)

Résumé - Détecter des formes de réponses aberrantes sur des échelles multi-réponses: Des indices non-paramétriques et un logiciel. La littérature de psychologie expérimentale comprend plusieurs indices de mesurer censés identifier des formes de réponses aberrantes. Cet article examine trois indices non-paramétriques bien connus, donne des exemples d'application et présente un logiciel qui calcule ces indices. **Caractérisation de réponses d'individus, Caractérisation appropriée d'individus. Erreur de répondant.**

Abstract. The psychometric literature contains many indices that are aimed at detecting aberrant c.q. deviant response patterns. This paper discusses three well-known nonparametric indices, gives an example of an application, and describes a computer program that calculates these indices. **Person Fit Research, Appropriateness Research, Respondent Error.**

INTRODUCTION

For measures consisting of a number of separate items, such as multi-item scales used in questionnaires, the usual procedure is to calculate a sum score for each individual who answered the questions. However, in addition to this sum score, the *score pattern* also provides information about the respondent. For example: suppose that a student gives the correct answer to eight out of ten questions during an examination. When the questions are ordered as to difficulty, one would expect that the student has missed the two most difficult questions. However, if the student has missed the two *easiest* questions, the response pattern is quite unusual, and we are justified to look deeper into the matter.

The psychometric literature contains many approaches toward analyzing and evaluating individual response patterns. This field of research is usually referred to as 'person-fit research.' Person fit research generally takes place in the context of psychological and educational testing (Meijer, 1994). Person fit research has developed a number of indices to identify persons who have unexpected or *aberrant* response patterns, either with respect to a scaling model or with respect to the other response patterns in the sample. When a respondent has an aberrant score pattern, one may wonder whether the sum score has the same interpretation as for other respondents in the sample (Van der Flier, 1980, chapter 1). Here, the sum score is the result of a very dissimilar score pattern, and whether it still is a valid measure of the underlying construct is questionable.

For example, assume that two students both have a score of five on a ten-item test. When the questions are ordered as to difficulty, we observe that one student has given the correct answer to the five easiest questions, and has missed the five most difficult questions. The other student has given the correct answer to the first two easy questions, and to the last three most difficult questions. For the student with the expected score pattern, the sum score of five is probably a good indicator of the ability. For the student with the unexpected pattern, the sum score is probably not a good indicator. The unexpected score pattern points toward more complicated interpretations, such as highly specific lacunae, or maybe simply cheating.

SOME EMPIRICAL APPLICATIONS OF PERSON FIT RESEARCH

Person fit research has developed several measures of person fit. These measures generally indicate *whether* someone has an unusual response pattern, not *why* such a pattern might occur. Some early empirical applications of person fit indices are Harnisch and Linn (1981), Tatsuoka (1985) and Tatsuoka and Tatsuoka (1983). Harnisch and Linn (1981) analyzed test data from 110 schools, and calculated person fit indices for the pupils. The fit indices were subsequently used to identify schools where aberrant response patterns were relatively frequent. These school effects could be attributed to differences in the school curriculum, such as paying attention to the metric system. Tatsuoka (1985) and Tatsuoka and Tatsuoka (1983) investigated the responses of pupils on a math test that required them to add and subtract positive and negative numbers. They also used person fit indices to detect aberrant response patterns. Analysis of these response patterns identified pupils that used an incorrect algorithm, or had difficulties with specific tasks, such as not understanding what the 'absolute value of a number' is.

Van der Flier (1980) used person fit indices in intercultural research. The central research problem was whether the scores on psychological tests are comparable for individuals that come from different cultures. In his first study, he analyzed the individual scores of Kenyan and Tanzanian children on intelligence tests that were presented in Kiswahili. In general, Tanzanian children understand Kiswahili better than Kenyan children. He found that the degree of aberrancy of the intelligence test scores correlated with the degree of command of Kiswahili; children that had a poor understanding of Kiswahili tended to have more aberrant test score patterns. Also, for Kenyan pupils with low aberrancy scores, the intelligence tests predicted their exam results better then the test results of pupils with high aberrancy scores. This indicates that high aberrancy scores coincide with test scores that are less valid. Van der Flier's second study investigated whether having western 'testing skills' would influence the test results. The response patterns of Kenyan pupils were compared with those of British pupils to construct a specific fit index: a 'western testing skill' index. For Kenyan pupils with few western testing skills, the intelligence tests underestimated their exam results. Also, the western testing skill index correlated as predicted with background variables such as the social-educational status of the parents.

Person fit indices have been applied to survey research by De Leeuw and Hox (1988), Van Tilburg and De Leeuw (1991), De Leeuw (1992), Meijer and De Leeuw (1993) and De Leeuw and Hox (1994). De Leeuw and Hox (1988) investigated the effect of successive data collection waves in mail surveys. A second or third data collection wave for the nonrespondents is a suitable procedure for raising survey response rates. However, it also implies more pressure on prospective respondents. De Leeuw and Hox (1988) showed that this does not affect the validity of the responses, as reflected by person fit indices on several multi-item scales.

Van Tilburg and De Leeuw (1991) performed secondary analyses on six surveys in the Netherlands. These surveys differ in the average person fit. Self-administered surveys resulted in fewer aberrant response patterns than face-to-face surveys. This is consistent with findings by De Leeuw (1992), who reported that a mail survey resulted in higher scale reliability, as indicated by classic psychometric indices, than a face-to-face or telephone survey.

Meijer and De Leeuw (1993) used person fit indices to subdivide their respondents in a group with many and a group with few aberrant responses. Respondents with many aberrant responses were significantly older and had less education than respondents with expected response patterns.

De Leeuw and Hox (1994) compared four well-known nonparametric person fit indices for several multi-item scales. They found that the various indices correlated highly among themselves, but had low correlations across different scales. Correlations of person fit indices with background variables were also low. They concluded that response aberrancy is not a fixed individual characteristic, but the result of an interaction between respondents and question characteristics. They suggest to use person fit indices to identify specific subgroups of respondents who may have problems with specific questions. Follow-up studies, for instance using cognitive laboratory methods (Forsyth and Lessler, 1991; Campanelli, 1997), can then be used to find out why these questions are difficult for that specific subgroup, and to suggest improvements for the questions.

SOME SPECIFIC PERSON FIT INDICES

Person fit indices are measures that indicate to what extend an individual response pattern is unusual or unexpected. Most indices have a high value when the response pattern is unexpected, which makes them, in fact, aberrancy indices. Three types of person fit indices can be distinguished: parametric indices, nonparametric indices, and group-based indices.

Parametric person fit indices are explicitly based on a parametric Item Response Theory (IRT). Parametric IRT indices model the responses to a multi-item scale by specific functions; they may or may not include assumptions about the population score distribution. Examples of parametric IRT indices are the one-parameter logistic model, which is better known as the Rasch model, and the three-parameter logistic model, which is also known as the Birnbaum model. For parametric person fit indices, it is necessary to specify a specific IRT model and to demonstrate that the model fits the data. Next, the model and the specific parameter estimates are used to calculate the probability of individual response patterns. If a specific response pattern has a low probability (conditional on the model and the parameter estimates), that response pattern is considered aberrant, and individuals with that response pattern are poorly scalable. For an overview of parametric person fit indices, see Kogut (1986) and Molenaar and Hoijtink (1990).

Nonparametric person fit indices are based on nonparametric IRT models, usually the Mokken model for monotone homogeneity. The assumptions of this model are explained by Mokken and Lewis (1982), and Meijer, Sijtsma and Smid (1990). Van der Flier (1980), and Sijtsma and Meijer (1992) have developed person fit indices based on this model. These indices assume that the Mokken model of monotone homogeneity holds for the empirical data, and calculate the probability of specific response patterns under that assumption. For an overview of nonparametric person fit indices, including the formulas and algorithms, we refer to Meijer (1994); we restrict ourselves to a general description.

According to Van der Flier (1980), a specific response pattern is aberrant, when it has a low probability, compared to other response patterns in a subsample with the same total score. To determine if a specific response pattern is aberrant, Van der Flier (1980) proposes to use the probability of that response pattern, given the total score, plus all response pattern probabilities that are lower than or equal to that probability. This procedure results in a measure that Van der Flier calls Q, which can be interpreted as a one-sided p-value that reflects the probability of finding that specific response pattern or one that is even more unusual. High values of Q indicate response patterns that are well within the expected range, while values of Q below 0.05 indicate responses that are aberrant at a significance level of alpha = 0.05. The advantage of Van der Flier's Q is that it has this convenient statistical interpretation. The disadvantage is that it implies calculation of all possible response pattern probabilities. This takes some computer time, and, with large multi-item scales, can easily exceed the computer capacity. To address this problem, Van der Flier proposes an index called U3. This index is not based on individual response pattern probabilities, but on the order of the item difficulties. Van der Flier (1980), and Meijer, Molenaar and Sijtsma (1994) demonstrate that even for fairly small scales, the index U3 has approximately a standard normal distribution. If a one-sided test is desired, a value for U3 exceeding 1.29 (the 90th percentile of the standard normal distribution) indicates an aberrant response pattern. The U3 index can be calculated much faster than the Q index. With small numbers of items (less than 20), U3 is probably inferior to Q, but this is not a major problem, because, with less than 20 items, it is quite possible to calculate the exact probability using the value of Q.

Sijtsma and Meijer (1992) take a different approach to nonparametric person fit under the Mokken model. The standard Mokken model (Mokken, 1970; Mokken and Lewis, 1982) uses Loevinger's H (Loevinger, 1947) to indicate how well a specific item fits in the scale. Sijtsma and Meijer propose to transpose the data matrix, giving persons the role of items and vice versa, and to calculate Loevinger's H for persons instead of items. This procedure is straightforward, but has the disadvantage that, with large numbers of persons, the computer capacity may be exceeded.

Group-based person fit indices rely on the empirical results of the total group that is analyzed. Group-based indices do not make the strict assumptions of various parametric and nonparametric IRT models. The criterion of aberrancy is the degree of similarity between a specific response pattern and the other response patterns in the data, for persons with comparable total scores. Response patterns that deviate too much from the typical response patterns in a specific group are considered aberrant. The departure point is the ideal Guttman-pattern. When the items are ordered as to difficulty (in a specific score-group), the ideal Guttman-pattern would be that an individual, who has a total score of n out of k items, would have answered correctly to the n easiest items, and incorrectly to all following k-n more difficult items. Both Harnisch and Linn (1981) and Meijer (1994) provide an extensive review of group-based person fit indices. Both reviews come to the same conclusion, that the 'modified caution index' C* is to be preferred. This index is based on Sato's caution index, which is zero for perfect Guttman patterns, but has no definite upper boundary. The modified caution index C*, proposed by Harnisch and Linn (1981), has the upper boundary of one when a response pattern is equal to the reverse of the ideal Guttman pattern. The modified caution index is an aberrancy index; higher values indicate more aberrant response patterns. There is no formal cut-off point; on the basis of simulations, Harnisch and Linn propose to interpret values of C* that exceed 0.3 (or, if one prefers to be more strict, exceeding 0.5) as an indication of an aberrant response process.

AN EMPIRICAL EXAMPLE

By way of example, we present two small simulation studies, following procedures developed by Meijer (1994). In our first study, a data set was generated with seven items and 1,500 cases. The simulation data were generated from a Rasch model with discrimination parameter equal to 2, equally spaced item difficulties ranging from -2 to +2, and a standard normal distribution for the latent trait. The resulting scale is reasonably reliable, with a Cronbach alpha of 0.73 and a Loevinger H of 0.85. We use Q, U3 and C* to classify the cases as normal versus aberrant. For Q and U3, we set the significance level at alpha = 0.10. For C*, we set the cut-off value at 0.30, following Harnisch and Linn (1981). Since all cases follow the parametric Rasch model, the correct classification would be 100% 'normal' cases. Actually, Q marks 11.2% of the cases as aberrant, U3 0%, and C* 11.8%. Since the nominal significance level is 0.10, we should expect about 10% false positives. The classification given by U3 is formally better than the other two classifications, but obviously, with only seven items, U3 does not follow a standard normal distribution.

The second simulation study generated a similar data set, but this time with 1,200 normal and 300 aberrant cases. For 150 cases, with a latent trait below zero, all item scores were replaced by randomly chosen values of 0/1 (with equal probability). These cases can be considered 'guessers'; they are cases with low capability who guess all responses. For another 150 cases, with a latent trait below zero, the values for the two most difficult items (item 6 and 7) were replaced by ones. These cases can be considered 'cheaters'; they have low capability but by some means obtain the correct answer for the two most difficult items. Thus, in this data set, 20% of the cases are actually aberrant response patterns. The psychometric properties of this data set are marginal; Cronbach's alpha is 0.54 and Loevinger's H is 0.30.

Van der Flier's Q marks 20.1% of the cases as aberrant. Only 3% of the normal cases are classified as aberrant, and 13% of the aberrant cases are classified as normal. The modified caution index, C*, performs about equally well; 17.5% of the cases are classified as aberrant. Using C*, 0.3% of the normals are classified as aberrant, and 13% of the aberrants classified normal. Again, U3 classifies zero percent of the cases as aberrant, which implies that 100% of the aberrant cases are classified as normal. Both Q and C* perform about equally well in classifying both guessers and cheaters.

In both simulation studies, the three aberrancy indices correlate all above 0.88. Thus, if the goal is to correlate an aberrancy measure with other respondent attributes, any of these measures can be used. If the goal is to classify respondents as 'normal' versus 'aberrant,' U3 is a poor choice. The indices Q and C* perform about equally well in this case.

CONCLUSIONS

Parametric IRT models come with strong assumptions, and in practice often do not show a satisfactory fit to many interesting data sets. In general, person fit indices that are based on less restrictive assumptions are preferable.

If the data follow the assumptions of monotone homogeneity of the Mokken model, and the number of items is not too large, then Van der Flier's Q is attractive, because it can be interpreted as a one-sided p-value. If the number of items is too large to allow the calculation of Q, Van der Flier's U3 can be used to approximate the p-value using a standard normal distribution. Simulation shows that with more than 20 items, the assumption that U3 follows a standard normal distribution is acceptable (Van der Flier, 1980; Meijer, Molenaar and Sijtsma, 1994).

If there are doubts about the assumption of monotone homogeneity, Harnisch and Linn's (1981) modified caution index C* may be used. This index uses the perfect Guttman pattern as a criterion, but uses the empirical response patterns of the total (score) group as a reference to assess the degree of aberrancy.

Educational and psychological testing generally use questionnaires with a large number of items. In sociological research and opinion surveys, usually very short multi-item scales are used (Heath and Martin, 1997). With long tests, both U3 and C* can be used, depending on the assumptions that are valid for the data. With short tests, Q or C* can be used, again depending on the validity of the scaling model. For an extensive review of the properties of various aberrancy indices, under different circumstances of test length and reliability, see Meijer (1994) and Meijer *et al.* (1994).

Given its modest assumptions, C* is probably the best overall choice.

THE PROGRAM ABERRANT

The program ABERRANT calculates Van der Flier's Q and U3 and Harnisch and Linn's C* for large samples of persons and a maximum of 70 items. It should be noted that the response patterns corresponding to a perfect or a zero sum score are unique and therefore can not be assigned a value for these indices. Such response patterns receive a missing value code of 9 in the fit indices. ABERRANT also calculates and outputs the sum score.

Input. The program expects a data file consisting of a respondent identification number, followed by dichotomous [0,1] item scores. The input is in free format (variables are separated by at least one space), which is what SPSS produces by default. Since calculating Q is time consuming, with the required computing time approximately doubling for each added item, the program will ask the user is Q is to be calculated.

Missing Data. An item score outside the range [0,1] is automatically interpreted as a missing value. In such cases, no aberrancy indices or sum scores can be calculated, and these are set to the missing value indicator 9 (the sum score is set to the missing value indicator 999). Response patterns that consist of all zeros, or all ones, also result in missing values for the aberrancy indices, but in these cases a sum score can be calculated.

Output. The program outputs a raw data file with the respondent identification number, the values for Q, U3 and C*, and the sum score, in standard format (F6.0,3F6.2,F6.0). These can be read into SPSS as a raw data file in free format. For Q, U3, and C*, the missing value code is 9, for the sum score 999. If the user has specified that Q should not be calculated, the output format is the same, with all values for Q set to missing.

Program. ABERRANT is an MSDOS program that is written in Turbo Pascal. It is available by writing to the second author at <hox@educ.uva.nl>. The standard version of ABERRANT assumes that a mathematical coprocessor is available (80287 and higher), but a version for PC's without a coprocessor is available on request. Both versions will run under Windows.

NOTE

The authors thank Rob Meijer and Klaas Sijtsma for their comments on earlier versions of this paper.

REFERENCES

Campanelli, P. (1997). "Testing Survey Questions: New Directions in Cognitive Interviewing". *Bulletin de Méthodologie Sociologique*, 55, 5-17.

De Leeuw, E.D. (1992). *Data Quality in Mail, Telephone, and Face to Face Surveys.* Amsterdam: TT-Publikaties.

De Leeuw, E.D. & Hox, J.J. (1988). "Artifacts in Mail Surveys". In: W.E. Saris & I.N. Gallhofer. *Sociometric Research, Volume 2 Data Analysis.* London: MacMillan, 61-73.

De Leeuw, E.D. & Hox, J.J. (1994). "Are Inconsistent Respondents Consistently Inconsistent? A Study of Several Nonparametric Person

Fit Indices". In: J.J. Hox & W. Jansen (eds). *Measurement Problems in Social and Behavioral Research.* Amsterdam: SCO-Kohnstamm Instituut, 67-88.

Forsyth, B.H. & Lessler, J.T. (1991). "Cognitive Laboratory Methods: A Taxonomy". In: P.P. Biemer *et al.* (eds). *Measurement Errors in Surveys.* New York, Wiley.

Harnisch, D.L. & Linn, R.L. (1981). "Analysis of Item Response Patterns: Questionable Test Data and Dissimilar Curriculum Practices". *Journal of Educational Measurement,* 18, 133-146.

Heath, A. & Martin, J. (1997). "Why Are There so Few Formal Measuring Instruments in Social and Political Research?". In: L. Lyberg *et al.* (eds.). *Survey Measurement and Process Quality.* New York: Wiley.

Kogut, J. (1986). *Review of IRT-based Indices for Detecting and Diagnosing Aberrant Response Patterns.* Enschede: Universiteit Twente, Toegepaste Onderwijskunde, Rapport 86-4.

Loevinger, J. (1947). "A Systematic Approach to the Construction and Evaluation of Tests of Ability". *Psychological Monographs,* 61, 4.

Meijer, R.R. (1994). *Nonparametric Person Fit Analysis.* Ph.D. Thesis, Vrije Universiteit, Amsterdam.

Meijer R. R. & De Leeuw, E.D. (1993). "Person Fit Indices in Survey Research: The Detection of Respondents with Unexpected Response Patterns". In J.H.L. Oud & R.A.W. van Blokland-Vogelesang (eds.), *Advances in Longitudinal and Multivariate Analysis in the Behavioral Sciences.* Nijmegen: ITS, chapter 17, 236-245.

Meijer, R.R., Sijtsma, K. & Smid, N.G. (1990). "Theoretical and Empirical Comparison of the Mokken and Rasch Approach to IRT". *Applied Psychological Measurement,* 14, 283-298.

Meijer, R.R., Molenaar, I.W. & Sijtsma, K. (1994). "The Influence of Test and Person Characteristics on Nonparametric Appropriateness Measurement". *Applied Psychological Measurement,* 18, 111-120.

Meijer, R.R. & Sijtsma, K. (1995). "Detection of Aberrant Item Score Patterns: A Review of Recent Developments". *Applied Measurement in Education,* 8, 261-272.

Mokken, R.J. (1970). *A Theory and Procedure of Scale Analysis.* 's Gravenhage, NL: Mouton & Co.

Mokken, R.J. & Lewis, C. (1982). "A Nonparametric Approach to the Analysis of Dichotomous Item Responses". *Applied Psychological Measurement*, 6, 417-430.

Molenaar, I.W. & Hoijtink, H. (1990). "The Many Null Distributions of Person Fit Indices". *Psychometrika*, 55, 75-106.

Sijtsma, K. & Meijer, R.R. (1992). "A Method for Investigating the Intersection of Item Response Functions in Mokken's Nonparametric IRT Model". *Applied Psychological Measurement*, 16, 149-157.

Tatsuoka, K.K. (1985). "A Probabilistic Model for Diagnosing Misconceptions by the Pattern Classification Approach". *Journal of Educational Statistics*, 10, 55-73.

Tatsuoka, K.K. & Tatsuoka, M.M. (1983). "Spotting Erroneous Rules of Operation by the Individual Consistency Index". *Journal of Educational Measurement*, 20, 221-230.

Van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties*. Lisse: Swets & Zeitlinger.

Van Tilburg T. & De Leeuw, E.D. (1991). "Stability of Scale Quality under Various Data Collection Procedures: A Mode Comparison on the 'De Jong-Gierveld Loneliness Scale'". *International Journal of Public Opinion Research*, 3, 69-85.
