Listening. Learning. Leading.®

Research Report
ETS RR–12-13

A Note on Explaining Away and Paradoxical Results in Multidimensional Item Response Theory

Peter W. van Rijn

Frank Rijmen

August 2012

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

A Note on Explaining Away and Paradoxical Results in Multidimensional Item Response Theory

Peter W. van Rijn and Frank Rijmen
ETS, Princeton, New Jersey

August 2012

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Associate Editor: Matthias von Davier

Reviewers: Sandip Sinharay and Frederic Robin

Abstract

Hooker and colleagues addressed a paradoxical situation that can arise in the application of multidimensional item response theory (MIRT) models to educational test data. We demonstrate that this MIRT paradox is an instance of the explaining-away phenomenon in Bayesian networks, and we attempt to enhance the understanding of MIRT models by placing the paradox in a broader statistical modeling perspective.

Key words: multidimensional IRT, paradoxical results, explaining away, Bayesian networks

Acknowledgments

The authors would like to thank Matthias von Davier, Shelby Haberman, and Bob Mislevy for helpful comments.

Hooker, Finkelman, and Schwartzman (2009) addressed a paradoxical situation that can arise in the application of multidimensional item response theory (MIRT) models to educational test data. The paradox boils down to the fact that a correct response on an additional item can lead to a lower estimate for one of the latent ability variables, whereas an incorrect response can lead to a higher estimate (Van der Linden, 2012). Hooker et al. (2009) argued that this is unfair to test takers. Various different appearances, generalizations, and implications of the paradox have been studied by numerous authors over the past few years (Finkelman, Hooker, & Wang, 2010; Hooker, 2010; Hooker & Finkelman, 2010; Jordan & Spiess, 2012; Van der Linden, 2012). The stated paradoxical situation is related to the explaining-away phenomenon in Bayesian networks (Pearl, 2009; Wellman & Henrion, 1993), which in statistics is known as Berkson's paradox (Berkson, 1946). In this report, we demonstrate that the MIRT paradox is an instance of this phenomenon, and we attempt to enhance the understanding of MIRT models by placing the paradox in a broader statistical modeling perspective, namely, that of graphical models and Bayesian networks (Mislevy, 1994; Pearl, 2009; Williamson, 2005). These frameworks provide a shorthand for the probabilistic relationships of interest and can help understand the properties of these relationships. We discuss a small number of MIRT modeling examples in these frameworks, illustrating the relation between the MIRT paradox and the explaining-away phenomenon, and we end with some concluding remarks.

1 Examples

In the following examples, we will adhere to parametric IRT in the framework of generalized nonlinear mixed models (Mellenbergh, 1994; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003), and we will make additional assumptions as needed; that is, we do not make assumptions about the types of items (continuous or discrete; dichotomous or polytomous), the types of latent variables (continuous or discrete), and the response functions (linear, normal, or logistic). We assume that both item response variables and latent variables are random and that item response variables can be observed, whereas latent variables cannot. (Because we make as few assumptions as possible, standard linear

factor models are included here as well.) An important assumption in both unidimensional and multidimensional IRT models is monotonicity. *Monotonicity* requires the probabilities for the item variables to be strictly increasing or decreasing in each latent variable, and MIRT models are monotone if and only if the latent variables are compensatory (Holland & Rosenbaum, 1986; Van der Linden, 2012). Strictly speaking, we do not need to make the monotonicity assumption, but then a unidimensional IRT model for which local independence holds can always be specified for a set of item variables (Suppes & Zanotti, 1981). Therefore we need to keep the assumption of monotonicity and will illustrate other assumptions, such as local independence, through the examples. In all our examples, we have chosen to use six items to keep things simple yet nontrivial. Furthermore, we assume that the first five items are already observed so that the sixth item is always the focal additional item that possibly creates the paradoxical situation.

Figure 1 displays a partially directed acyclic graph (DAG) of a MIRT model with two latent variables θ_1 and θ_2 and six item response variables $X_1, X_2, \ldots, X_6$. (It is called *partially directed* because not all the lines in the graph have arrowheads. A partial DAG is also referred to as a *chain graph*.) This model is said to be of simple structure, also referred to as a between-item two-dimensional IRT model, because every item response variable is linked to a single latent variable only. In the graph, the nodes correspond to random variables, and the directed edges represent conditional dependency relations. An advantage of using graphical models is that there is a correspondence between the property of separation of the nodes in the graph and conditional independence of the random variables in the statistical model. For example, the path $X_1 \leftarrow \theta_1 \rightarrow X_2$ in Figure 1 illustrates an instance of so-called d-separation (Pearl, 2009, pp. 16–17); that is, the only path from X_1 to X_2 runs through θ_1, and the arrows do not meet head to head at θ_1. The fact that X_1 and X_2 are d-separated in the graph implies that they are conditionally independent given θ_1. We can generalize this to all six items in the example, and obtain the familiar IRT assumption of local independence: the joint probability of $X_1, X_2, \ldots, X_6$ is conditional on θ_1 and θ_2 can be written as a simple product: $\Pr(X_1, X_2, \ldots, X_6|\theta_1, \theta_2) = \prod_{j=1}^{3} \Pr(X_j|\theta_1) \prod_{j=4}^{6} \Pr(X_j|\theta_2)$. Because of the correspondence between d-separation and conditional independence, it is

possible to determine all conditional independence relations that are entailed solely by working with the graph. Now, the MIRT paradox revolves around the beliefs about θ_1 and θ_2 in different situations. In describing the paradox, Hooker et al. (2009) always seemed to condition implicitly on $X_1, X_2, \ldots, X_5$. Keeping this in mind, the MIRT paradox cannot arise for the model in Figure 1 because the only path between θ_1 and θ_2 is the undirected edge; that is, conditional on $X_1, X_2, \ldots, X_5$, the additional observation of X_6 does not affect the belief about θ_1 in an unexpected manner.

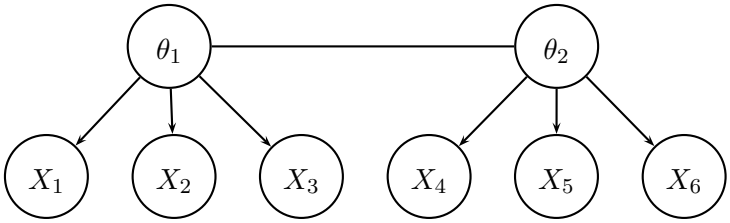

Figure 1. **Partially directed acyclic graph of two-dimensional item response theory model with between-item multidimensionality.**

Figure 2 shows the DAG of a two-dimensional IRT model for six items with so-called within-item multidimensionality for items 3 and 4. In this figure, the paths $\theta_1 \rightarrow X_3 \leftarrow \theta_2$ and $\theta_1 \rightarrow X_4 \leftarrow \theta_2$ are so-called inverted forks and contain the first and foremost step of explaining what happens in the MIRT paradox. These paths between θ_1 and θ_2 are not blocked by X_3 and X_4 because the edges on these paths meet head to head. Therefore θ_1 and θ_2 are not d-separated by X_3 and X_4, and conditional independence between θ_1 and θ_2 given X_3 and X_4 is not implied. We note that this kind of conditional independence is different from that typically used in IRT because we condition here on observed variables instead of on unobserved variables. Now, even if θ_1 and θ_2 are independent a priori, they become dependent when we condition on $X_1, \ldots, X_5$. Furthermore, the observation of X_6 can affect the belief about θ_1 in an unanticipated fashion. This at first sight counterintuitive phenomenon is called the *explaining-away effect*. We refrain from giving substantive

examples to be concise and because intuitive examples of this phenomenon are described by many authors (e.g., Berkson, 1946; Bishop, 2006, p. 378; Hooker & Finkelman, 2010, p. 251; Pearl, 2009, p. 17).

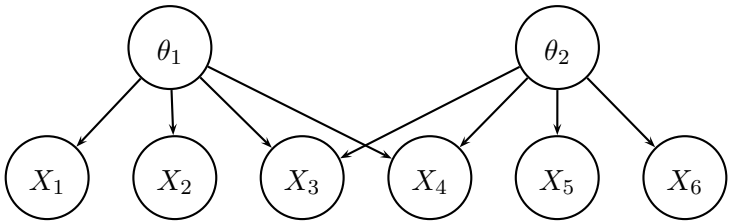

Figure 2. **Directed acyclic graph of two-dimensional item response theory model with within-item structure.**

We emphasize that this explaining-away phenomenon can arise as long as there is at least one inverted fork on the paths between θ_1 and θ_2 through $X_1, X_2, \ldots, X_5$ that does not depend on the particular relation of θ_1 and θ_2 with X_6. We illustrate this by two other instances of the phenomenon. The first case is illustrated in Figure 3, in which the focal sixth variable is not an item response but the variable gender, where gender is related to θ_2. Obviously, observing gender changes the belief about θ_2, but the belief about θ_1 can be affected in an unexpected manner owing to the inverted forks. Again, this dependency can arise when θ_1 and θ_2 are a priori independent and when θ_1 is unrelated to gender (as is the case in Figure 3). This example is particularly interesting because many applications of multidimensional IRT models with background variables are found in large-scale assessments such as the Programme for International Student Assessment (PISA; Adams, Wilson, & Wang, 1997) and the National Assessment of Educational Progress (NAEP; Mislevy, 1985). (However, we note that the current MIRT models in PISA and NAEP have a between-item structure, as in Figure 1.) A second instance can be constructed when we relate gender to an item response variable instead of to a latent variable. This situation is given in Figure 4, where gender-related differential item functioning appears on the fifth item. Observing gender affects the belief about θ_2 through X_5 as well as the belief about θ_1

because of the inverted forks. To reiterate, paradoxical results in all these instances are not to be attributed to the focal sixth variable but to the inverted forks in other parts of the model.

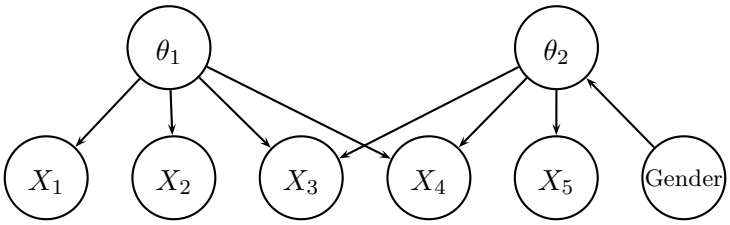

Figure 3. **Directed acyclic graph of two-dimensional item response theory model with within-item structure and relation between gender and θ_2.**

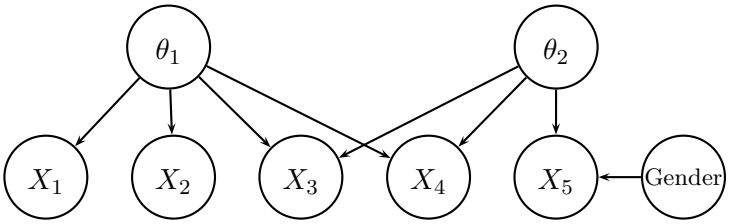

Figure 4. **Directed acyclic graph of two-dimensional item response theory model with within-item structure and gender-related differential item functioning for X_5.**

Hooker and Finkelman (2010) considered the MIRT paradox in models for item bundles. They focused on two models: the bifactor model and the testlet model. In the bifactor model, every item loads on a general dimension and on an item bundle dimension. Hooker and Finkelman discussed two cases, one in which all latent variables are assumed to be independent and one in which the item bundle dimensions are correlated. Independent latent variables are typically assumed to identify the bifactor model, which is the situation

that we consider. An example of the bifactor model is represented in a DAG in Figure 5. Hooker and Finkelman consider a result to be paradoxical if answering an additional item (X_6) correctly results in a lower estimate for the general ability (θ_1) than when the additional item is answered incorrectly. From Figure 5, it follows that θ_1 and θ_3 are not d-separated, that is, there are paths between θ_1 and θ_3 that contain an inverted fork (in fact, all paths do). Hence the explaining-away phenomenon can occur, and paradoxical results are possible for this bifactor model. Hooker and Finkelman (2010) derived mathematically the specific conditions under which paradoxical results occur for the more general bifactor model. From their mathematical derivations, it follows that paradoxical results are not possible when the loadings of the bifactor model are restricted according to the so-called testlet model (a testlet model is a restricted bifactor model; see Rijmen, 2010). The fact that paradoxical results cannot occur for the testlet model (with independent nuisance dimensions) can be shown directly by looking at the corresponding DAG, alleviating the need for mathematical derivations. First, one should realize that the testlet model is a Schmid–Leiman transformed second-order model (see, e.g., Yung, Thissen, & McLeod, 1999). Then, the conditional independence relations can be observed from the DAG of the equivalent second-order model, which is presented in Figure 6. In this figure, it is easily seen that θ_1 and θ_3 are always dependent because the path from θ_1 to θ_3 has a directed edge. However, θ_1 is independent from X_4, X_5, and X_6 is conditional on θ_3; that is, conditional on θ_3, the observation of X_6 does not change the belief about θ_1 in an unexpected manner. Therefore, as long as monotonicity holds, paradoxical results cannot occur in this case.

2 Concluding Remarks

We have shown that the MIRT paradox utilized by Hooker et al. (2009) is an instance of the explaining-away phenomenon. Specifically, the so-called inverted fork in the path between latent variables is the main cause of the phenomenon. In many of the MIRT paradox papers, intuitions are built up from an educational measurement perspective, which causes the result to be surprising. However, we made use of the frameworks of graphical models and Bayesian networks in which this phenomenon is well established. We chose

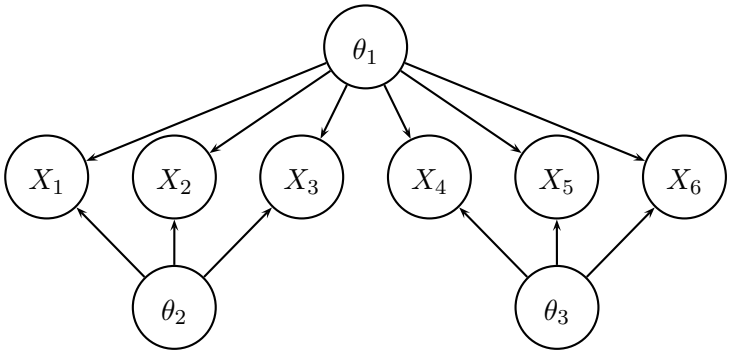

Figure 5. **Directed acyclic graph of bifactor three-dimensional item response theory model.**

these frameworks because the conditional dependencies between the variables in a specific model can be derived directly from its graph, independent of different parameterizations and link functions.

The work of Hooker et al. (2009) is nevertheless to be lauded because they described the exact mechanics of the paradox in MIRT in great detail. We disagree, however, with the somewhat pessimistic conclusions of Jordan and Spiess (2012) and Van der Linden (2012) on the usefulness of MIRT models. The MIRT paradox is a general statistical paradox that holds for many models with multiple competing explanatory variables and is accepted in many contexts other than psychometrics such as biostatistics and artificial intelligence. We find that the issue of test fairness raised by Hooker et al. (2009) and Jordan and Spiess (2012) results from confounding different views on the purpose of tests. For example, Holland (1994) distinguished between tests as contests and tests as measurement. The contest view can result in a firm belief that more items correct should result in a higher score, a feature that nevertheless pertains to relatively few IRT models (Van der Linden, 2012). In the measurement view, model selection is perhaps the most important issue so that test-based inferences are sound. A third view on tests, raised by Mislevy (1994), suggests that tests can be used as sources of information for evidentiary reasoning about students, for example, as in models for cognitive diagnosis. Preventing paradoxical results

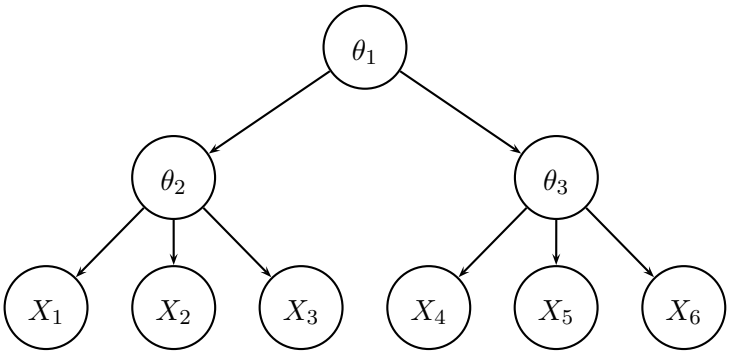

Figure 6. **Directed acyclic graph of second-order (or testlet) three-dimensional item response theory model.**

might be relevant in the contest perspective on tests, but we argue that it is less relevant in the latter two perspectives on the purposes of educational tests.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Berkson, J. (1946). Limitations of the application of fourfold tables to hospital data. *Biometrics Bulletin, 2*, 47–53.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York: Springer.

Finkelman, M., Hooker, G., & Wang, J. (2010). Prevalence and magnitude of paradoxical results in multidimensional item response theory. *Journal of Educational and Behavioral Statistics, 35*, 744–761.

Holland, P. W. (1994). Measurements or contests? Comments on Zwick, Bond and Allen/Donoghue. In *Proceedings of the Social Statistics Section of the American Statistical Association* (pp. 27–29). Alexandria, VA: American Statistical Association.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics, 14*, 1523–1543.

Hooker, G. (2010). On separable tests, correlated priors, and paradoxical results in multidimensional item response theory. *Psychometrika, 75*, 694-707.

Hooker, G., & Finkelman, M. (2010). Paradoxical results and item bundles. *Psychometrika, 75*, 249–271.

Hooker, G., Finkelman, M., & Schwartzman, A. (2009). Paradoxical results in multidimensional item response theory. *Psychometrika*, 74, 419–442.

Jordan, P., & Spiess, M. (2012). Generalizations of paradoxical results in multidimensional item response theory. *Psychometrika, 77*, 127–152.

Lord, F. M. (1962). Cutting scores and errors of measurement. *Psychometrika, 27*, 19–30.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 300–307.

Mislevy, R. J. (1985). Estimating latent group effects. *Journal of the American Statistical Association, 80*, 993–997.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika,*

59, 439–483.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York: Cambridge University Press.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*, 361–372.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*, 185–205.

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese, 48*, 191-199.

Van der Linden, W. J. (2012). On compensation in multidimensional response modeling. *Psychometrika, 77*, 21–30.

Wellman, M. P., & Henrion, M. (1993). Explaining "explaining away." *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15*, 287–292.

Williamson, J. (2005). *Bayesian nets and causality.* Oxford: Oxford University Press.

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*, 113–128.