British Journal of Mathematical and Statistical Psychology (2014)

The British Psychological Society

www.wileyonlinelibrary.com

On the explaining-away phenomenon in multivariate latent variable models

Peter van Rijn[1]* and Frank Rijmen[2]

[1]ETS Global, Amsterdam, The Netherlands
[2]CTB/McGraw-Hill, Monterey, California, USA

Many probabilistic models for psychological and educational measurements contain latent variables. Well-known examples are factor analysis, item response theory, and latent class model families. We discuss what is referred to as the 'explaining-away' phenomenon in the context of such latent variable models. This phenomenon can occur when multiple latent variables are related to the same observed variable, and can elicit seemingly counterintuitive conditional dependencies between latent variables given observed variables. We illustrate the implications of explaining away for a number of well-known latent variable models by using both theoretical and real data examples.

1. Introduction

Many statistical models for psychological and educational measurements make use of latent variables. Well-known model families are factor-analytic models, item response theory (IRT) models, structural equation models, and latent class models. Typically, such models contain conditional independence assumptions; for example, the observed variables are independent given the latent variables. In this paper, we discuss an important dependence relationship that can arise in the context of multiple latent variables, which is generally known as the 'explaining-away' phenomenon (Pearl, 1988). This phenomenon can occur when two or more latent variables are related to the same observed variables, and amounts to a change in the dependence between two latent variables when certain variables are observed. In its most basic form, a negative dependence arises while the latent variables are initially independent. This is referred to as explaining away, and can be understood as follows. Suppose that there are two possible (unobserved) causes for a given event. Observing the event increases the likelihood for each of the causes. However, upon observing other evidence for the first cause to be present, the other cause becomes less likely as the events have already been explained by the likely presence of the first cause. Even though intuitively easily understood with discrete latent variables (presence or absence of causes), a similar phenomenon can occur with continuous latent variables or a combination of discrete and continuous latent variables. More formally, the increase in the posterior probability of one latent variable can reduce the need for other latent variables (Wellman & Henrion, 1993). We discuss several latent structures that can give rise to this phenomenon, and focus on both theoretical and practical implications. The explaining-away phenomenon is not new, because in essence it is an instance of Berkson's

*Correspondence should be addressed to Peter W. van Rijn, ETS Global, Strawinskylaan 929, 1077 XX Amsterdam, The Netherlands (email: pvanrijn@etsglobal.org).

DOI:10.1111/bmsp.12046

paradox (Berkson, 1946). However, recent work in the context of multidimensional IRT by Hooker, Finkelman, and Schwartzman (2009) has spawned a series of papers, but we argue that the phenomenon is relevant in the more general context of latent variable models. Our main message is that explaining away in a latent variable model can be shown to be related to general model properties, which can be pointed out by making use of the framework of graphical models. We will illustrate that this avoids the need to having to prove many seemingly different theorems for specific latent variable models.

In our discussion of the explaining-away phenomenon, we make use of the frameworks of generalized linear mixed models (McCulloch & Searle, 2001) and graphical models (Lauritzen, 1996). Pearl (1988) argued for a probabilistic approach in dealing with uncertainty in modeling a particular problem. He claimed that brute force manipulation of probabilistic models could neither become technically feasible nor substantively acceptable. The path ahead was modularity through the imposition or assumption of meaningful conditional independence relations. A natural way to do this is by using graphical models or Bayesian networks. Pearl (1988) realized that graphs not only provide an attractive way for communicating complex structures, but also form the basis for efficient algorithms for propagating evidence. The roots of such an argument date back to statistical mechanics (Gibbs, 1902) and genetics (Wright, 1921), but can also be found in log-linear models (Darroch, Lauritzen, & Speed, 1980).

Bollen and Bauldry (2011) argue that typically in factor analysis, classical test theory, IRT, and structural equation models the assumption is that the measures are *effect* indicators in contrast to *causal* indicators, composite indicators, and covariates. That is, in these models, latent variables are typically interpreted as causes, and observed variables as effects. Wellman and Henrion (1993) point out the asymmetry in the reasoning from cause to effect (causal reasoning) and from effect to cause (evidential reasoning). For example, the assumption of local independence that is common in many unidimensional IRT models is an instance of causal reasoning. That is, the observed variables are independent conditional on the latent variable. However, the explaining-away phenomenon is an instance of evidential reasoning. That is, the dependence between latent variables changes upon conditioning on observed variables. Although there are some exceptions (Holland & Rosenbaum, 1986; Mislevy, 1994), little attention has been paid to the latter type of reasoning in the literature on latent variable modelling, and our aim is to fill this gap.

The explaining-away phenomenon has, however, been discussed in the context of multidimensional item response theory (MIRT) models by van Rijn and Rijmen (2012). They illustrate that the so-called paradoxical results in MIRT (Hooker *et al.*, 2009) are an instance of explaining away. However, Hooker *et al.* (2009) use a very different, more empirical point of departure. That is, they observed that a correct response on an additional item can lead to a lower estimate for one of the latent ability variables, while an incorrect response can lead to a higher estimate. This occurs even though the probability correct is a monotonically increasing function for each of the latent variables. It can even be the case that in a practical setting where a pass/fail decision is made, there are two students with the same responses for all but the last item, where the student with the last item correct fails and the student with the last item incorrect passes (Hooker *et al.*, 2009, p. 419). Statistical properties related to this issue have been addressed from a test fairness perspective in a multitude of papers (Hooker, 2010; Hooker & Finkelman, 2010; Hooker *et al.*, 2009; Jordan & Spiess, 2012; van der Linden, 2012), and it is claimed that such a result is paradoxical.

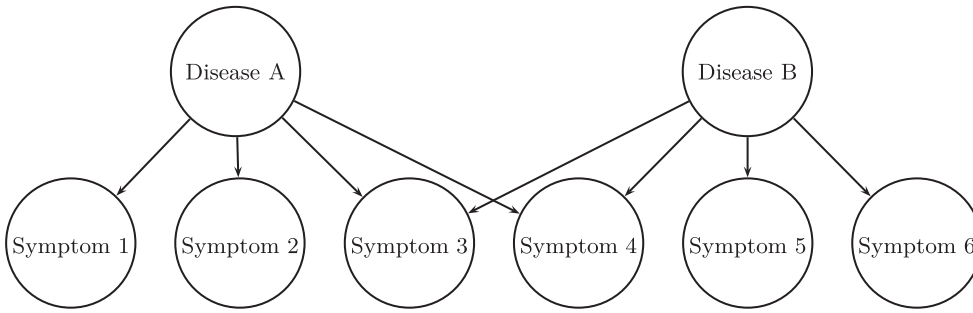

Figure 1. Graph for example of two diseases and six symptoms.

We observe that the MIRT paradox seems counterintuitive from a practical fairness perspective, whereas many examples of the explaining-away phenomenon are quite compelling. Consider the example of a set of symptoms that can be caused by two diseases (see Figure 1).[1] Although the diseases can be independent and not all symptoms have to be related to both diseases, observing the symptoms provides information about both diseases. First, upon observing symptoms 3 and 4, both disease A and disease B become more likely, since both symptoms are caused by either. However, upon observing either one of the other four symptoms, the belief in one disease is increased, but at the same time the belief in the other is reduced. For example, upon observing symptom 6 conditional on the presence or absence of the other five symptoms, the belief in disease B increases. However, since the other five symptoms have not changed, this means that the belief in disease A has to be adjusted downwards. So, the explaining-away effect renders a negative dependence between the two diseases even though they are a priori independent.

First, we discuss relevant aspects of latent variable models, such as conditional independence, monotonicity, and total positivity. In the following section, we introduce the framework of graphical models and give a precise definition of the explaining-away phenomenon. The reason for using this framework is that many latent variable models can be specified as graphical models. This operation has a number of amenities, for example conditional independence relationships can be determined directly from the graph. In the remainder of the paper, we discuss theoretical implications of the explaining-away phenomenon for a number of specific latent variables models and two empirical examples.

2. Assumptions of latent variable models

First, we introduce some necessary notation. We denote latent variables by Θ (e.g., abilities), observed dependent variables by Y (e.g., items), and observed independent variables by X (e.g., covariates). Realizations of these variables are denoted in lower case. Vectors and matrices are written in bold. Latent and observed variables can be continuous or discrete. Individuals are numbered $i = 1, 2, \ldots, n$, observed independent variables $j = 1, 2, \ldots, m$, and latent variables $k = 1, \ldots, d$.

[1] This example can be seen as a highly simplified version of the quick medical reference model, which consists of approximately 600 diseases and approximately 4,000 symptoms (see, for example, Jaakkola & Jordan, 1999).

Second, we introduce two main assumptions in latent variable models following Holland and Rosenbaum (1986). For continuous observations, we let the conditional distribution of a set of m observed dependent variables $\mathbf{Y} = \mathbf{y}$ given a set of d latent variables $\boldsymbol{\Theta} = \boldsymbol{\theta}$ be given by

$$F(\mathbf{y}|\boldsymbol{\theta}) = F(y_1, \ldots, y_m|\boldsymbol{\theta}) = \Pr(Y_1 \leq y_1, \ldots, Y_m \leq y_m|\boldsymbol{\theta}). \tag{1}$$

Then the assumption of *conditional independence* can be given by

$$F(\mathbf{y}|\boldsymbol{\theta}) = \prod_{j=1}^{m} F(y_j|\boldsymbol{\theta}). \tag{2}$$

The same holds when we consider the density $f(\mathbf{y}|\theta)$:

$$f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{j=1}^{m} f(y_j|\boldsymbol{\theta}). \tag{3}$$

For discrete observations, the above density is replaced by a probability function $\Pr(\mathbf{y}|\boldsymbol{\theta})$. Suppes and Zanotti (1981) have shown that conditional independence can always be obtained, although the resulting latent variable model might be futile. A meaningful addition then is the assumption of *monotonicity* which holds if all $1 - F(y_j|\boldsymbol{\theta})$ are non-decreasing in each latent dimension for all values y_j. Monotonicity is equivalent to the following monotone likelihood ratio property (Milgrom, 1981):

$$\frac{f(y_j|\boldsymbol{\theta})}{f(y_j|\boldsymbol{\theta}')} \geq \frac{f(y_j'|\boldsymbol{\theta})}{f(y_j'|\boldsymbol{\theta}')}, \tag{4}$$

for each j, $y_j > y_j'$, and all $\boldsymbol{\theta} \geq \boldsymbol{\theta}'$, that is, $\theta_k \geq \theta_k'$ for $k = 1, 2, \ldots, d$. Monotonicity is implied if multivariate total positivity of order 2 (MTP_2) holds for the conditional densities $f(y_j|\boldsymbol{\theta})$ (Karlin & Rinott, 1980). MTP_2 for a random vector $\mathbf{Y}$ holds if

$$f(\max(\mathbf{y}, \mathbf{y}^*))f(\min(\mathbf{y}, \mathbf{y}^*)) \geq f(\mathbf{y})f(\mathbf{y}^*). \tag{5}$$

Note that there is no ordering on $\mathbf{y}$ and $\mathbf{y}^*$. Although the theory of total positivity (Karlin, 1968) is relevant for our topic and we return to it later on, a full discussion is beyond our present requirements. For a more elaborate discussion of total positivity and other forms of positive dependence in latent variable models and some interrelations, see Holland and Rosenbaum (1986), and, in a MIRT context, Jordan and Spiess (2012). In addition, Joe (1997) discusses relevant dependence concepts.

Our initial focus is on latent variable models that fit in the framework of generalized linear mixed models (Hedeker, 2005). If we denote the mean for person i of Y_{ij} by μ_{ij}, we can use (see Rijmen, Tuerlinckx, Boeck, & Kuppens, 2003, eq. 4)

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{v}_{ij}'\boldsymbol{\beta} + \mathbf{w}_{ij}'\boldsymbol{\theta}_i, \tag{6}$$

where $g(.)$ is the link function, η_{ij} is the linear predictor, $\mathbf{v}_{ij}$ is a p-dimensional covariate vector for the fixed effects, and $\mathbf{w}_{ij}$ is a d-dimensional covariate vector for the random

effects. In the generalized linear mixed model framework, it is assumed that the distribution of Y_i is of the exponential family type. Also, the distribution for the random effects $\boldsymbol{\theta}_i$ is often assumed to be multivariate normal with mean zero and variance–covariance matrix $\boldsymbol{\Sigma}$ (Hedeker, 2005). Since we also want to accommodate models that contain multiplications of the parameters (such as factor loadings or item discriminations and a latent variable), we extend the framework to non-linear mixed models. The mean for such models can be denoted by (Skrondal & Rabe-Hesketh, 2004)

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{v}'_{ij}\boldsymbol{\beta} + \boldsymbol{\alpha}'\mathbf{W}_{ij}\boldsymbol{\theta}_i, \tag{7}$$

where $\boldsymbol{\alpha}$ is a q-dimensional vector for the random effects and $\mathbf{W}_{ij}$ is a $q \times d$ design matrix for the random effects. For example, a two-factor model for the graph in Figure 1 can be specified as follows:

$$E(\mathbf{Y}_i|\boldsymbol{\theta}_i) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} + \begin{bmatrix} \alpha_{11} & 0 \\ \alpha_{21} & 0 \\ \alpha_{31} & \alpha_{32} \\ \alpha_{41} & \alpha_{42} \\ 0 & \alpha_{52} \\ 0 & \alpha_{62} \end{bmatrix} \begin{bmatrix} \theta_{i1} \\ \theta_{i2} \end{bmatrix}. \tag{8}$$

For this example, MTP_2 for each $f(y_j|\boldsymbol{\theta})$ holds if all non-zero elements of $\boldsymbol{\alpha}$ are positive, and thus monotonicity holds. If, in addition, $\boldsymbol{\theta}$ is normal with non-negative covariance, then $f(\mathbf{Y})$ is MTP_2 (see, for example, Holland & Rosenbaum, 1986, Theorem 7). The analogous two-dimensional compensatory IRT model can be written in the same way by replacing the left-hand side of equation (7) with logit $(\Pr(\mathbf{Y}_i = 1|\boldsymbol{\theta}_i))$. For MTP_2, the same conditions should hold as before. Note that non-negativity of the slopes leads to log-negative second derivatives with respect to $\boldsymbol{\theta}$ of the conditional probabilities $\Pr(Y_{ij} = 1|\boldsymbol{\theta})$. This feature plays an important role in the proofs of paradoxical results in MIRT (Hooker *et al.*, 2009).

The structure of the product $\boldsymbol{\alpha}'\mathbf{W}_{ij}$ plays an important role in our analysis and examples. Since different terminologies are used to describe the latent structures in factor analysis and IRT, we briefly discuss both (see also McDonald, 2000). In factor-analytic terms, a multivariate latent variable model is generally referred to as having simple structure if each observed variable measures only one latent variable (Thurstone, 1947).[2] In MIRT, this would be referred to as between-item multidimensionality (see, for example, Adams, Wilson, & Wang, 1997). A multivariate factor model in which observed variables measure multiple traits is said to be of complex structure, which, in MIRT, would be described as within-item multidimensionality. Special cases of models with complex structure are the bifactor model (Holzinger & Swineford, 1937) and the higher-order factor model (Schmid & Leiman, 1957). The higher-order model can be subdivided into classes of higher-order models dependent on how many latent layers make up the hierarchy; that is, a model with two latent layers is referred to as a second-order model, etc. The bifactor model carries the same name in the MIRT framework (Gibbons & Hedeker, 1992), but the testlet model (Wainer, Bradlow, & Wang, 2007) and model for item bundles

[2] Thurstone's original definition of simple structure is less strict than our usage. We use this stricter form in order to align simple structure in factor analysis with the term between-item multidimensionality used in IRT (Rijmen & de Boeck, 2005). The correct factor-analytic term would be perfect cluster configuration.

(Hooker & Finkelman, 2010; Rosenbaum, 1988; Wilson & Adams, 1995) used in IRT are in fact instances of a second-order model. Formal relations between these models are described by Yung, Thissen and McLeod (1999) and Rijmen (2010).

3. Explaining away

We start this section with some graphical modelling concepts and conventions (for a more technical treatment, see Lauritzen, 1996). Nomenclature varies somewhat in the literature; we stick to that of Bishop (2006, Ch. 8). A graph is a diagram that consists of nodes and links. In the context of graphical (probabilistic) models, the nodes represent random variables and the links represent the dependence relationships between the nodes. The links can be directed or undirected. Directed graphical models have graphs in which all links are directed. An important class of directed graphs are graphs that contain no cycles (i.e., one cannot start and end with the same node by following the direction of the links). Models that can be represented by directed acyclic graphs (DAGs) are also referred to as Bayesian networks. Undirected graphical models have graphs that contain only undirected links and are referred to as Markov random fields. Graphs that contain both directed and undirected links are referred to as chain graphs. The latent variable models we consider are either DAGs or chain graphs. If there is a directed link from one node to another, then the originating node is referred to as a parent and the receiving node is a child. Finally, a positive link indicates a positive dependence between the nodes (e.g., a positive correlation), and a negative link indicates a negative one.

An important concept in DAGs that implies specific conditional dependence relationships is *d*-separation (Pearl, 1988 2009). Formally, two disjoint sets of variables A and B in a DAG are *d*-separated by a third disjoint set of variables C if and only if the path from A to B through C contains a chain ($A \rightarrow C \rightarrow B$) or fork ($A \leftarrow C \rightarrow B$), or an inverted fork ($A \rightarrow M \leftarrow B$) such that neither M nor any of its children are in C. Now, consider the unidimensional latent variable model in Figure 2. The path $Y_1 \leftarrow \Theta_1 \rightarrow Y_2$ illustrates an instance of *d*-separation. That is, the only path from Y_1 to Y_2 runs through Θ_1, and the arrows do not meet head-to-head (or do not collide) at Θ_1. The fact that Y_1 and Y_2 are *d*-separated by Θ_1 in the graph implies that they are conditionally independent given θ_1. We can generalize this to all six observed variables in the example, so that the joint probability of $y_1, y_2, \ldots, y_6$ conditional on θ_1 can be written as a simple product: $\Pr(y_1, y_2, \ldots, y_6|\theta_1, \theta_2) = \prod_{j=1}^{6} \Pr(y_j|\theta_1)$. Because *d*-separation implies conditional independence, it is possible to determine all conditional independence relations that are entailed solely by working with the (directed acyclic) graph. A main benefit of using a DAG is that the joint distribution can often be factorized into much more manageable conditional distributions by making use of the implied conditional independence relationships. We emphasize that the conditional dependence and independence relations that can be derived from the graph hold for all probability distributions.

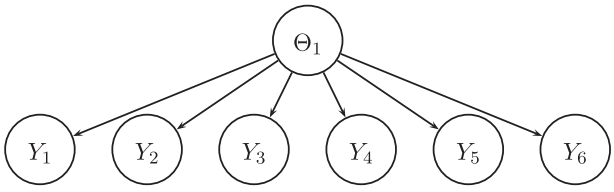

Figure 2. Directed acyclic graph of unidimensional latent variable model.

To conclude, the graph can offer insight into whether or not two (sets of) variables are conditionally independent given a third (set of) variable(s). That is, we can rely on the graphical representation to partition the latent variable models considered in this paper into two categories: those in which prior (conditional) independence of latent variables is preserved after observing the observed variables, and those in which the prior (conditional) independence is not preserved. If the prior independence is preserved, the explaining-away phenomenon cannot occur. This means that we can define a class of latent variable models for which the explaining-away phenomenon cannot occur, irrespective of the distributional assumptions and the parametrization of the model. For the second class of models, an analysis of the graph alone will not be sufficient to determine whether observing the responses will result in the explaining-away phenomenon (i.e., a negative dependence between latent variables). For that, a more detailed analysis of the model is needed.

Now the occurrence of the explaining-away phenomenon for two latent variables can be defined following Wellman and Henrion (1993) by the following theorem:

Theorem 1. (Explaining Away). Assume that conditional independence and monotonicity hold for $f(\mathbf{y}|\theta_1, \theta_2)$, and that θ_1 and θ_2 are marginally independent. Furthermore, let θ_1 and θ_2 be parents of $\mathbf{y}$. Then θ_1 and θ_2 become negatively dependent upon observing $\mathbf{y}$ if and only if

$$\frac{f(\mathbf{y}|\theta_1, \theta_2)}{f(\mathbf{y}|\theta_1, \theta_2')} \leq \frac{f(\mathbf{y}|\theta_1', \theta_2)}{f(\mathbf{y}|\theta_1', \theta_2')}, \tag{9}$$

for all $\theta_1 > \theta_1'$, $\theta_2 > \theta_2'$.

Note that even though other nodes can be part of the graph, the theorem only concerns nodes with multiple parents. The proof is given by Wellman and Henrion (1993, p. 291), but we repeat it here for the sake of completeness and because there are some conceptual and notational differences.

Proof. If θ_1 and θ_2 remained independent upon observing $\mathbf{y}$, we could write $f(\theta_1|\theta_2, \mathbf{y}) = f(\theta_1|\mathbf{y})$. Therefore, we are interested in the distribution of θ_1 given θ_2 and $\mathbf{y}$, which is given by

$$f(\theta_1|\theta_2, \mathbf{y}) = \frac{f(\mathbf{y}|\theta_1, \theta_2)\, f(\theta_1|\theta_2)}{f(\mathbf{y}|\theta_2)}. \tag{10}$$

We define θ_1 and θ_2 to be negatively dependent upon observing $\mathbf{y}$ if and only if $f(\theta_1|\theta_2, \mathbf{y})$ has the following monotone likelihood ratio property:

$$\frac{f(\theta_1|\theta_2, \mathbf{y})}{f(\theta_1|\theta_2', \mathbf{y})} \leq \frac{f(\theta_1'|\theta_2, \mathbf{y})}{f(\theta_1'|\theta_2', \mathbf{y})}. \tag{11}$$

If we expand this using Bayes' theorem, we get

$$\frac{f(\mathbf{y}|\theta_1,\theta_2)f(\theta_1|\theta_2)f(\mathbf{y}|\theta_2')}{f(\mathbf{y}|\theta_1,\theta_2')f(\theta_1|\theta_2')f(\mathbf{y}|\theta_2)} \leq \frac{f(\mathbf{y}|\theta_1',\theta_2)f(\theta_1'|\theta_2)f(\mathbf{y}|\theta_2')}{f(\mathbf{y}|\theta_1',\theta_2')f(\theta_1'|\theta_2')f(\mathbf{y}|\theta_2)}, \tag{12}$$

which simplifies to

$$\frac{f(\mathbf{y}|\theta_1,\theta_2)f(\theta_1|\theta_2)}{f(\mathbf{y}|\theta_1,\theta_2')f(\theta_1|\theta_2')} \leq \frac{f(\mathbf{y}|\theta_1',\theta_2)f(\theta_1'|\theta_2)}{f(\mathbf{y}|\theta_1',\theta_2')f(\theta_1'|\theta_2')}. \tag{13}$$

In addition, θ_1 and θ_2 are independent, so we can write

$$\frac{f(\mathbf{y}|\theta_1,\theta_2)}{f(\mathbf{y}|\theta_1,\theta_2')} \leq \frac{f(\mathbf{y}|\theta_1',\theta_2)}{f(\mathbf{y}|\theta_1',\theta_2')}. \tag{14}$$

□

Wellman and Henrion (1993, p. 289) state that explaining away requires that 'the proportional increase in the probability of $\mathbf{y}$ on raising θ_2 is smaller for higher values of θ_1'.

A theorem for explaining away can also be derived for the case where Θ_1 and Θ_2 are a priori dependent (Wellman & Henrion, 1993, Theorem 2). This dependence between the parents should, however, be of the same type as the parent–child dependence. If the dependence is negative, explaining away strengthens the negative dependence. If the dependence is positive, the relationship becomes ambiguous. That is, explaining away decreases the dependence, but it also depends on the strength of the a priori dependence and the observations if the sign is flipped or not. This case is particularly important for psychometrics because many factor and MIRT models allow correlations between the latent variables. Examples of such models are the confirmatory five-factor model of personality (Borkenau & Ostendorf, 1990) and the operational MIRT model in the US National Assessment of Educational Progress (NAEP; von Davier, Sinharay, Oranje, & Beaton, 2007). The case of a priori correlated $\boldsymbol{\theta}$ is studied in the context of MIRT by Hooker *et al.* (2009) and Hooker (2010). We will elaborate further on correlated latent variables in the upcoming modelling examples.

We now illustrate the explaining-away phenomenon with a simple example using the DAG in Figure 3. We can derive from the graph that the latent variables are independent, and that the observed variables are independent conditional on the latent variables. Let us further assume that the latent variables are continuous, the observed variables are dichotomous, and a two-dimensional compensatory two-parameter logistic IRT model holds (Birnbaum, 1968; Reckase, 2009). The model for Figure 3 can then be given by

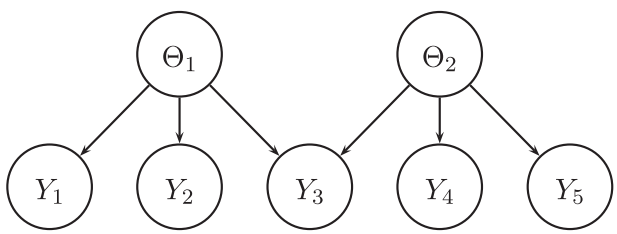

Figure 3. Directed graph of two-dimensional latent variable model with complex structure.

$$\text{logit}(\Pr(\mathbf{Y}_i = 1|\boldsymbol{\theta}_i)) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} \begin{bmatrix} \alpha_{11} & 0 \\ \alpha_{21} & 0 \\ \alpha_{31} & \alpha_{32} \\ 0 & \alpha_{42} \\ 0 & \alpha_{52} \end{bmatrix} \begin{bmatrix} \theta_{i1} \\ \theta_{i2} \end{bmatrix}. \tag{15}$$

Without loss of generality, let all item intercepts be 0 and all non-zero item slopes be 1. The above explaining-away theorem concerns the likelihoods of response patterns. Using the graph in Figure 3, we can simply obtain the likelihood (Bishop, 2006, p. 362)

$$\begin{aligned} L(\theta_1, \theta_2|\mathbf{y}) = f(\mathbf{y}|\theta_1, \theta_2) &= p(y_1, y_2, \ldots, y_5|\theta_1, \theta_2) \\ &= p(y_1|\theta_1)p(y_2|\theta_1)p(y_3|\theta_1, \theta_2)p(y_4|\theta_2)p(y_5|\theta_2). \end{aligned} \tag{16}$$

We only need the part where both parents are involved. In order to prove that explaining away can occur, we need to show that

$$\frac{p(y_3|\theta_1, \theta_2)}{p(y_3|\theta_1, \theta_2')} \leq \frac{p(y_3|\theta_1', \theta_2)}{p(y_3|\theta_1', \theta_2')}. \tag{17}$$

This reduces to showing that

$$\exp(\theta_1 + \theta_2') + \exp(\theta_1' + \theta_2) \leq \exp(\theta_1' + \theta_2') + \exp(\theta_1 + \theta_2). \tag{18}$$

If we let $\delta_1 = \theta_1 - \theta_1'$ and $\delta_2 = \theta_2 - \theta_2'$, we can write

$$\exp(\delta_1) + \exp(\delta_2) \leq 1 + \exp(\delta_1 + \delta_2). \tag{19}$$

Because the exponential function is convex, the above inequality is easily verified, which proves that explaining away can occur for this model.

We could continue with proving explaining-away theorems for many specific latent variable models without reference to the framework of graphical models, but the graphical formalism provides some benefits and insights (Smyth, Heckerman, & Jordan, 1997). For example, it is relatively straightforward to see that in Figure 3, the inverted fork $\Theta_1 \rightarrow Y_3 \leftarrow \Theta_2$ enables explaining away. It can be shown that if a DAG contains inverted forks, explaining away can occur. However, this condition is not sufficient, which is why the theorem is needed. The theorem can be generalized to models with more than two latent variables, but this is not pursued here.

4. Dependent latent variables

We now discuss explaining away for an important class of models, that is, models in which the latent variables are dependent. As a straightforward example, we discuss a model in which the latent variables are dependent and each observed variable is linked to only one latent variable. As noted, this is referred to as simple structure and it is often used in psychological and educational measurement. In dealing with simple structure models with correlated latent variables, it is important to realize that independence of the latent variables can be easily achieved. For example, the Cholesky decomposition can be applied

to the covariance matrix of the latent variables (Bartholomew, 1984). If the inverse transformation is applied to the factor loadings or item discrimination, then the resulting model is observationally equivalent. Note, however, that with this transformation simple structure is lost.

Consider the two-dimensional latent variable model with simple structure as shown in Figure 4. To ease the presentation, we write vectors of observed variables in bold. We use a directed link between the latent variables, but this does necessarily not imply a causal relation (e.g., reversing the direction of all links between latent variables would not change the dependency structure of the graph). From the graph, we can find the joint probability $p(\mathbf{y}_1|\theta_1)p(\mathbf{y}_2|\theta_1)p(\theta_2|\theta_1)p(\theta_1)$. However, in most MIRT and FA cases, we only use $p(\mathbf{y}_1|\theta_1)p(\mathbf{y}_2|\theta_1)p(\theta_1, \theta_2)$. This illustrates that the direction of the link between Θ_1 and Θ_2 is arbitrary, because if it were reversed then the last two terms would be replaced by $p(\theta_1|\theta_2)p(\theta_2)$, and both are equal to $p(\theta_1, \theta_2)$. If we follow the steps of the proof of Theorem 1, it can easily be seen that explaining away reduces to showing that

$$\frac{p(\theta_1, \theta_2)}{p(\theta_1, \theta_2')} \leq \frac{p(\theta_1', \theta_2)}{p(\theta_1', \theta_2')},$$

for all $\theta_1 > \theta_1'$, $\theta_2 > \theta_2'$. This is not true if $p(\theta_1, \theta_2)$ is, for example, bivariate normal, because the normal distribution has the monotone likelihood ratio property in the opposite direction. However, as was shown by Hooker (2010) for both MIRT and factor models, if there are three correlated latent variables in a simple structure model, explaining away can occur.

Hooker (2010) used a three-dimensional simple structure model as an example. In order to understand the difference between the simple structure model with two and the model with three latent variables, we transform the directed graph into an undirected graph. This process is called moralization (Bishop, 2006, p. 391) and consists of two steps. First, the nodes that have a common child are connected, and then all directed links are replaced by undirected links. In order to illustrate the arbitrariness of the directions of the links between the latent variables, Figure 5 shows three different DAGs of the same simple structure model and the associated moral graph, which is the same for all three DAGs.

Hooker (2010) provided an example of a three-dimensional simple structure MIRT model to show that negative posterior dependencies (i.e., explaining away/paradoxical results) can arise. In the example, the following covariance matrix for the normally distributed, zero-mean latent variables is used:

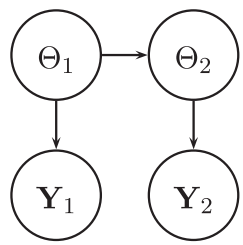

Figure 4. Directed acyclic graphs for two-dimensional latent variable model with simple structure.

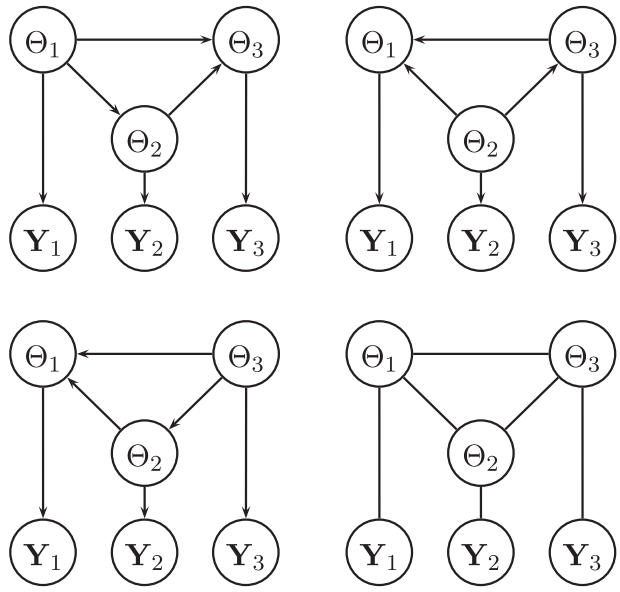

Figure 5. Three equivalent directed acyclic graphs and moral graph (bottom right) for three-dimensional latent variable model with simple structure.

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.00 & 0.80 & 0.26 \\ 0.80 & 1.00 & 0.60 \\ 0.26 & 0.60 & 1.00 \end{bmatrix}.$$

It can easily be verified that this matrix is not MTP_2, because its inverse contains positive off-diagonal elements (see Karlin & Rinott, 1980, p. 480). In addition, any two-dimensional covariance matrix with a positive covariance is MTP_2. This explains the fact that explaining away can occur for simple structure models with three or more dimensions, but not for two-dimensional models. Hooker (2010, p. 697) made use of the term $(\boldsymbol{\Lambda}'\boldsymbol{\Lambda} + \boldsymbol{\Sigma}^{-1})^{-1}$ that is used in maximum a posteriori estimation of $\boldsymbol{\theta}$ in MIRT and in computing factor scores in factor analysis, where $\boldsymbol{\Lambda}$ is the matrix with item discriminations or factor loadings, respectively. He showed that if this term has negative entries so-called paradoxical results can occur (i.e., by lowering an observed variable, one can get an increase in one of the latent variables). Note that $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}$ is a diagonal matrix for simple structure models and therefore the result is equivalent to $\boldsymbol{\Sigma}$ being MTP_2. That is, explaining away cannot occur in simple structure models if all free elements of $\boldsymbol{\Lambda}$ are positive and $\boldsymbol{\Sigma}$ is MTP_2.

In the approach in Hooker (2010), it is assumed that the correlations between the latent variables are specified a priori whereas they are typically estimated from the data (in most cases, the variances are fixed at one for model identification). This means that the maximum a posteriori and expected a posteriori estimates of the latent variables are empirical Bayes estimators. So, the occurrence of explaining away is an empirical question. A simple check on the estimated matrix will be sufficient to see if explaining away can occur. In addition, if one is interested in particular response patterns, one can look for negative posterior covariances of the latent variables. For example, for the three-dimensional simple structure for the seven-item example in Hooker (2010, Table 1),

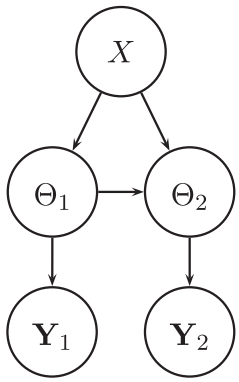

Figure 6. Directed acyclic graph of two-dimensional latent variable model with simple structure and covariate.

the element in $\boldsymbol{\Sigma}^{-1}$ associated with Θ_1 and Θ_3 is positive (1.2087), so MTP_2 does not hold and explaining away can occur. More specifically, for 116 of the 132 possible response patterns with seven dichotomous items, the posterior covariance between Θ_1 and Θ_3 is negative.

As a final example for this section, consider the graph in Figure 6 in which a covariate for the latent variables is introduced. This example is particularly interesting, because many applications of MIRT models with background variables are found in large-scale assessments such as the Programme for International Student Assessment (PISA; Adams *et al.*, 1997) and the NAEP (Mislevy, 1985; von Davier *et al.*, 2007). In addition, many structural equation models have a similar structure. Again, the direction between the latent variables is arbitrary, and we can derive the following joint probability of interest from the graph:

$$p(\mathbf{y}_1|\theta_1, x)p(\mathbf{y}_2|\theta_1, x)p(\theta_1, \theta_2|x). \tag{20}$$

Note that the distribution of X does not play a role in this joint probability, so that the results from before apply. However, if the covariate operates on the level of observed variables, for example as in the case of differential item functioning, explaining away can arise (see, for example, van Rijn & Rijmen, 2012).

5. Other latent structures

We aim to further discuss the implications of the explaining-away phenomenon for latent variable modelling by means of two historical examples of latent structures and two more recent examples. We start with the bifactor structure as proposed by Holzinger and Swineford (1937). This latent structure has received increased attention in the MIRT literature over the past two decades because of innovations in its estimation and its theoretical appeal (Gibbons & Hedeker, 1992; Reise, 2012; Rijmen, 2010). The DAG in Figure 7 is the structure that was hypothesized by Holzinger and Swineford (1937, Table 1). They applied the bifactor model to a set of tests on spatial ability, mental speed, motor speed, and verbal ability.

For explaining away between θ_1 and θ_2, for example, we have to show that

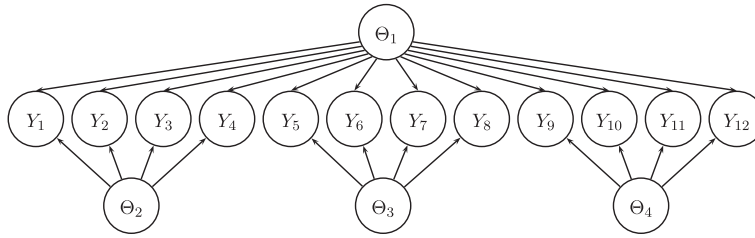

Figure 7. Directed acyclic graph of bifactor model as hypothesized in Holzinger and Swineford (1937).

$$\frac{f(y_1|\theta_1,\theta_2)}{f(y_1|\theta_1,\theta_2')} \leq \frac{f(y_1|\theta_1',\theta_2)}{f(y_1|\theta_1',\theta_2')}. \tag{21}$$

Now if we assume normal distributions and factor loadings equal to 1, and again use $\delta_1 = \theta_1 - \theta_1'$ and $\delta_2 = \theta_2 - \theta_2'$, this reduces to

$$\exp(\theta_1\theta_2' - \theta_1\theta_2) \leq \exp(\theta_1'\theta_2' - \theta_1'\theta_2) \tag{22}$$

$$\exp(-\delta_1\delta_2) \leq 1. \tag{23}$$

The above inequality is strict since $\delta_1 > 0$ and $\delta_2 > 0$. Hence, explaining away can occur in the bifactor model. This is indicated by the inverted forks in Figure 7. Of course, the same holds for bifactor IRT models (Gibbons & Hedeker, 1992), which has been shown by Hooker and Finkelman (2010). Reise (2012) recently argued that bifactor modelling is a good approach for psychometric modelling when a strong common factor is expected, but where multidimensionality arises due to groups of items from different subdomains. For example, in writing assessment, a bifactor IRT model with a general writing dimension and genre-specific dimensions could be an adequate structure for the writing construct. Typically, the factors in such a model are uncorrelated for identification purposes. However, since explaining away can occur, the latent variables become negatively dependent upon observing $\mathbf{y}$. As noted, this can be checked by inspecting posterior covariances.

The second historical example is the hierarchical or higher-order latent structure. Figure 8 shows a DAG of a third-order factor model as hypothesized by Schmid and Leiman (1957). In this case, the graphical formalism is very convenient, because the graph does not contain inverted forks and therefore explaining away cannot occur. It can be derived that explaining away cannot occur for models of this type, regardless of the order, and regardless of the number of orthogonal factors at the highest order. That is, no inverted forks are introduced by adding orders to the model. Note that a second-order model is a restricted bifactor model in both the factor-analytic case (Yung *et al.*, 1999) and the IRT case (Rijmen, 2010). When the dimensionality of $\boldsymbol{\Theta}$ is 4 or more, the second-order model has fewer parameters than the simple structure model.

As a final example, Figure 9 shows the DAG and moral graph of a speed and accuracy model for responses to test items as discussed by van der Linden (2007). The item responses are denoted by $\mathbf{Y}_1$, for which an IRT model is assumed, and the associated response times by $\mathbf{Y}_2$, for which a lognormal model is selected. van der Linden (2007) adds

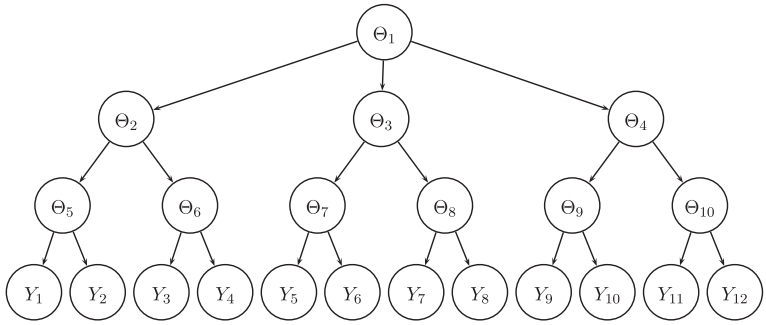

Figure 8. Directed acyclic graph of higher-order factor model as hypothesized in Schmid and Leiman (1957).

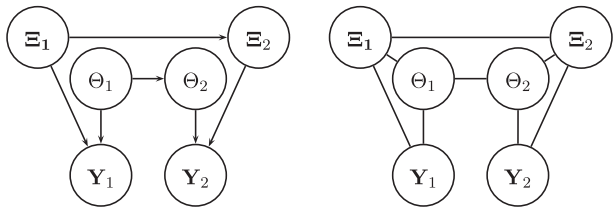

Figure 9. Directed acyclic (left) and moral (right) graph of speed and accuracy model with simple structure by van der Linden (2007).

an extra layer to the model due to a Bayesian approach to model estimation, so two sets of additional latent variables (item parameters) are introduced: one set for the item responses ($\boldsymbol{\Xi}_1$) and one for the response times ($\boldsymbol{\Xi}_2$). Now, again, all dependence relations in the model can be derived from the graph in Figure 9, and the joint distribution can be written as

$$p(\mathbf{y}_1|\theta_1, \boldsymbol{\xi}_1)p(\mathbf{y}_2|\theta_2, \boldsymbol{\xi}_2)p(\theta_1, \theta_2)p(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2).$$

Note that the direction of the link between the latent variables is arbitrary. The structure of this model is quite complex, so we first consider the case where Θ_1 and Θ_2 are independent. Now the question is whether Θ_1 and Θ_2 are *d*-separated by $\mathbf{Y}_1$ and $\mathbf{Y}_2$. The answer is no, because the path from Θ_1 to Θ_2 contains an inverted fork (in fact, two). If the latent item variables $\boldsymbol{\Xi}_1$ and $\boldsymbol{\Xi}_2$ are independent, then Θ_1 and Θ_2 are *d*-separated by $\mathbf{Y}_1$ and $\mathbf{Y}_2$. However, this is not the case in this model, since van der Linden (2007) explicitly allows $\boldsymbol{\Xi}_1$ and $\boldsymbol{\Xi}_2$ to be correlated. So, even if the two latent person variables are independent and are related to separate sets of observed variables, a negative dependence can be rendered between them upon observing $\mathbf{Y}_1$ and $\mathbf{Y}_2$. Again, this can be checked empirically by inspecting the posterior covariance matrix. When Θ_1 and Θ_2 are dependent, explaining away can also occur by the same reasoning (see also van der Linden, 2007, p. 296). Note that the phenomenon would have an interesting interpretation: conditional on a set of item responses, observing a lower response time for one item would result in a higher estimate for the latent variable associated with speed, but a lower estimate for the latent variable associated with accuracy.

We could continue with many other examples, but we only briefly mention latent class models, that is, models in which both the observed and latent variables are discrete. The latent class model in all its forms, including, for example, cognitive diagnostic models, can also be studied in the framework of graphical models (see, for example, Díez & Druzdzel, 2006; von Davier & Haberman, 2014). However, these models are only monotonic in the probabilities for observed variables by imposing parameter constraints (Henson, Templin, & Willse, 2009; van Onna, 2002). Such constraints are important in both model identification and parameter estimation. Therefore, a more detailed description would be needed in order to study the mechanics of the explaining-away phenomenon for such models, and this is beyond our present scope.

6. Empirical examples

6.1. Example 1: Factor analysis

In this section, we discuss explaining away in a factor-analytic context. We start with some notational conventions in factor analysis. We consider the factor model

$$\begin{aligned}\mathbf{Y} &= \boldsymbol{\Lambda}\boldsymbol{\Theta} + \boldsymbol{\varepsilon},\\ \boldsymbol{\Sigma} &= \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi},\end{aligned}$$

where $\boldsymbol{\Lambda}$ is the factor loading matrix, the covariance matrix of $\boldsymbol{\Theta}$ is denoted by $\boldsymbol{\Phi}$ and the covariance matrix of $\boldsymbol{\varepsilon}$ is given by $\boldsymbol{\Psi}$. We refer to a classic data set of Holzinger and Swineford as discussed, for example, in Jöreskog (1969). The data set consists of nine subtest scores from an intelligence test. Jöreskog (1969) fitted a series of factor models, and we refitted a selection of these models in R using the lavaan package (Rosseel, 2012).[3] For example, a three-dimensional simple structure model with correlated factors was fitted (Jöreskog, 1969, Table 1d). The estimate of the latent covariance matrix ($\boldsymbol{\Phi}$) with standard errors in parentheses for this model is given by

$$\begin{bmatrix} 1.00(-) & & \\ 0.46(0.06) & 1.00(-) & \\ 0.47(0.07) & 0.28(0.07) & 1.00(-) \end{bmatrix}.$$

This matrix is MTP_2, that is, its inverse has all positive off-diagonal elements, and explaining away does not occur. This can also be checked by computing the covariance matrix of the latent variable estimates, that is, the covariance matrix of the factor scores. There are many different methods for calculating factor scores (see, for example, Lawley & Maxwell, 1971, Ch. 8). In the normal case, the posterior of $\boldsymbol{\theta}$ given $\mathbf{y}$ is (Bartholomew, 1981, eq. 5)

$$\boldsymbol{\theta}|\mathbf{y} \sim \text{N}\Big(\left(\boldsymbol{\Phi}^{-1} + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}), \left(\boldsymbol{\Phi}^{-1} + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)^{-1}\Big).$$

Note that the posterior mean coincides with the typical regression factor scores. Hooker (2010, p. 697) notes that paradoxical results (explaining away) and negative posterior

[3] The data set comes with the lavaan package, so our analysis can be easily repeated (our code is available upon request).

correlation can occur if the posterior covariance matrix has negative off-diagonal elements. This is not the case in our example, where the posterior covariance matrix is given by

$$\begin{bmatrix} 0.28 & & \\ 0.02 & 0.11 & \\ 0.04 & 0.01 & 0.28 \end{bmatrix}.$$

The phenomenon that the covariance matrix of factor score estimates can be quite different from $\boldsymbol{\Phi}$ has long been known in factor analysis. Anderson and Rubin (1956) discuss factor score estimates that are so-called covariance-preserving. That is, these factor scores are obtained under the constraint that their covariance matrix is equal to $\boldsymbol{\Phi}$. There are several ways to accomplish this (see Neudecker, 2004). For example, Krijnen, Wansbeek, and ten Berge (1996) discuss a method in which the determinant of the mean squared error matrix $E\left((\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\right)$ is minimized. This estimator is given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{K}} = \boldsymbol{\Phi}^{\frac{1}{2}}\left(\boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{+}\boldsymbol{\Phi}^{\frac{1}{2}}\right)^{-\frac{1}{2}}\boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{+}\left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}\right),$$

where the + superscript indicates the Moore–Penrose inverse. Note, however, that the posterior covariance matrix is not the same as the covariance of the factor scores.

The best-fitting structure is the model as shown in Figure 10 (Jöreskog, 1969, Table 1g). Note that the structure is both bifactor and complex, since subtests 8 and 9 have loadings on all three factors. The estimated factor loadings and error variances (with standard errors in parentheses) for this model are given by

$$\begin{bmatrix} 0.47(0.07) & 0.68(0.08) & - \\ 0.22(0.07) & 0.51(0.08) & - \\ 0.19(0.07) & 0.72(0.08) & - \\ 0.99(0.06) & - & - \\ 1.10(0.06) & - & - \\ 0.91(0.05) & - & - \\ 0.17(0.07) & - & 0.74(0.08) \\ 0.17(0.06) & 0.25(0.06) & 0.68(0.08) \\ 0.26(0.06) & 0.46(0.06) & 0.44(0.06) \end{bmatrix} \text{ and } \begin{bmatrix} 0.68(0.09) \\ 1.08(0.10) \\ 0.72(0.10) \\ 0.36(0.05) \\ 0.45(0.06) \\ 0.36(0.04) \\ 0.60(0.11) \\ 0.47(0.08) \\ 0.54(0.06) \end{bmatrix}.$$

The covariance matrix of the three factors is the identity matrix. The posterior covariance matrix of $\boldsymbol{\theta}$ conditional on $\mathbf{y}$ for this model is then found to be

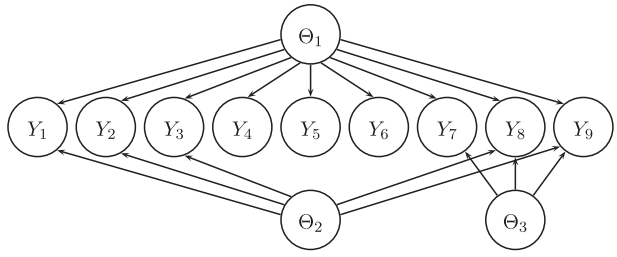

Figure 10. Directed acyclic graph of bifactor model for nine intelligence subtests as fitted by Jöreskog (1969).

$$\begin{bmatrix} 0.11 & & \\ -0.03 & 0.34 & \\ -0.01 & -0.07 & 0.32 \end{bmatrix}.$$

The negative off-diagonal values indicate that explaining away can occur in this model, while all factor loadings are positive and the factors are independent. We will illustrate this further by inspecting the first case of the data, which contains the following scores:

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
3.33	7.75	0.38	2.33	5.75	1.29	3.39	5.75	6.36

The factor scores (using the posterior mean) for these data are −0.13, −0.56, and 0.20 for the three factors, respectively. If we raise the score on subtest 5 from 5.75 to 6.75 (e.g., meaning one additional item correct), then the factor scores change to 0.14, −0.64, and 0.16, respectively. The decrease in the scores for factors 2 and 3 would be called a paradoxical result in the context of MIRT, because the observed score for subtest 5 increases but the factor scores decrease while all factor loadings are positive and the factors are independent. This is exactly the explaining-away mechanism in action: the increase in the general factor explains away factors 2 and 3 conditional on the associated observed subtest scores.

6.2. *Example 2: MIRT*

In our second empirical example, we make use of data from a writing assessment as described in van Rijn, Deane, Rijmen, and Bennett (2014).[4] There were four different assessments, one for each of four different writing purposes: recommendation writing, informational writing, argumentative writing, and literary analysis. A total of 1693 eighth-grade students in the United States took two of the four assessments in a counterbalanced incomplete design. The total number of items is 74, and a mix of dichotomous and polytomous items was used.

Table 1 shows the dimensionality, number of parameters, negative log-likelihood, Akaike's information criterion (AIC), and Bayesian information criterion (BIC) for three different models. The models are multivariate extensions of the two-parameter logistic model (2PLM) for dichotomous items and the generalized partial credit model (GPCM) for polytomous items. The four models have a unidimensional structure, simple structure, second-order structure, and bifactor structure, respectively. The models were estimated

Table 1. Comparative fit indices of IRT models for four writing assessments

Model	Dimensions	Parameters	−Log-likelihood	AIC	BIC
2PLM/GPCM	1	208	43,659.4	87,735	88,865
2PLM/GPCM simple structure	4	214	43,368.2	87,168	88,327
2PLM/GPCM second-order	5	212	43,371.9	87,164	88,321
2PLM/GPCM bifactor	5	282	43,173.0	86,910	88,443

[4] Other empirical and numerical examples in the context of MIRT can be found in Hooker *et al.* (2009), Hooker (2010), Hooker and Finkelman (2010), Jordan and Spiess (2012), and van der Linden (2012).

Table 2. Estimated correlations between dimensions for four-dimensional 2PLM/GPCM with simple structure

	1	2	3	4
1. Recommendation	–			
2. Informational	.81	–		
3. Argumentative	.88	.73	–	
4. Literary analysis	.82	.83	.84	–

with marginal maximum likelihood, using a program for item response analysis developed by Haberman (2013). The AIC prefers the bifactor model, while the BIC prefers the second-order model (although the difference with the simple structure model is small).

The correlations between the dimensions of the simple structure model are given in Table 2. Two correlations stand out: the correlation between informational and argumentative writing is lower than the others (.73), and the correlation between recommendation and argumentative writing is higher than the others (.88). It is easily verified that this matrix is not MTP_2, so explaining away can occur. However, none of the observed response patterns leads to a negative posterior covariance of $\boldsymbol{\theta}$. This is not so strange, because only a fraction of the possible response patterns is typically observed. Explaining away cannot occur for the second-order model, but it can occur for the bifactor model. This is confirmed by inspecting posterior covariances of $\boldsymbol{\theta}$ for both models. However, for the second-order model, one has to make sure which $\boldsymbol{\theta}$ is estimated (see Rijmen, 2010). The second-order model can be written as

$$\begin{aligned} g(\pi_{ij}) &= a_{jg}\theta^*_{ik} + b_j, \\ \theta^*_{ik} &= a_{kg}\theta_{ig} + \theta_{ik}, \end{aligned}$$

where θ_k are specific factors and θ_g is the general factor. If the θ_{ik} are estimated instead of the θ^*_{ik} (which is the case with the software we used), then the posterior covariances need to be transformed in order to get those from the graph of a second-order model. These posterior covariances are given by

$$\begin{aligned} \text{Cov}(\tilde{\theta}_{ig}, \tilde{\theta}^*_{ik}) &= a_{kg}\text{Var}(\tilde{\theta}_{ig}) + \text{Cov}(\tilde{\theta}_{ig}, \tilde{\theta}_{ik}), \\ \text{Cov}(\tilde{\theta}^*_{ik}, \tilde{\theta}^*_{ik'}) &= a_{kg}a_{k'g}\text{Var}(\tilde{\theta}_{ig}) + a_{k'g}\text{Cov}(\tilde{\theta}_{ig}, \tilde{\theta}_{ik}) + a_{kg}\text{Cov}(\tilde{\theta}_{ig}, \tilde{\theta}_{ik'}) + \text{Cov}(\tilde{\theta}_{ik}, \tilde{\theta}_{ik'}). \end{aligned}$$

For the second-order model in our example, we found negative covariances for $\text{Cov}(\tilde{\theta}_{ig}, \tilde{\theta}_{ik})$, but positive covariances for $\text{Cov}(\tilde{\theta}_{ig}, \tilde{\theta}^*_{ik})$. Negative posterior covariances were observed for the bifactor model, both between general and specific factors and between mutual specific factors. All results are in line with the theoretical expectations of the explaining-away phenomenon for the fitted models.

7. Discussion

We have discussed the explaining-away phenomenon in a variety of latent variable models. Although it has been studied recently mostly for IRT models from a test fairness perspective, the phenomenon is much more general. An important example from clinical psychology is differential diagnosis, which effectively exploits the explaining-away

phenomenon in a systematic way to find out which of the potential diagnoses receives the most support from the data. We made use of the framework of graphical models in which explaining away is well established (Pearl, 1988; Wellman & Henrion, 1993). We chose this framework because the conditional dependencies between the variables in a specific model can be derived directly from its graph, independent of different distributions, parametrizations and link functions. To reiterate, we do not need the framework of graphical models in order to show that explaining away can occur for certain latent variable models. However, the graphical formalism provides benefits for inference and insights (Smyth *et al.*, 1997), and our aim was to bring these benefits to bear in order to provide some insight into evidential reasoning in multivariate latent variable models. In order to make these insights concrete, we used some well-known latent structures as well as empirical examples to illustrate the workings of the explaining-away phenomenon.

If multivariate latent variable models are applied to scores on educational test items, we believe it is important to distinguish between tests of maximum performance (see, for example, Holland, 1994) and tests of typical performance (Mellenbergh, 2011) in discussing explaining away (or, equivalently, paradoxical results). That is, the balance between statistical optimality and social acceptability of test-based decisions for these two types of educational testing can be quite different. Even within each type, a different balance might need to be struck based on the stakes involved. Reckase and Luo (in press) argue however that the term 'paradoxical results' does not seem to bode well for estimation, whereas in fact it is good estimation. Nevertheless, the social acceptability of a scoring procedure might be deemed more important than good estimation in a specific context.

We focused our discussion on latent variables as random effects. If they are treated as fixed effects, maximum likelihood estimation can be carried out . In the context of MIRT, a special case that would deserve some attention is the so-called weighted maximum likelihood (WML) estimator (Warm, 1989). With a simple structure model, explaining away can occur with random effects, but not with fixed effects. This is, however, not addressed in the recent generalization of the WML estimator to MIRT (Wang, 2014), and a strong candidate for future study. Another such topic is related to the issue of treating model parameters as estimated instead of fixed. This can have an impact on using the posterior covariance matrix of the latent variables as a check on explaining away in practice (especially for simple structure models). Methods are available to account for the error that is due to variation in the estimation of model parameters (see, for example, Hoshino & Shigemasu, 2008).

Finally, another interesting topic for further research is how dependence relations might arise in evidential reasoning in conjunctive models with discrete latent variables (Junker & Sijtsma, 2001; Maris, 1999). These models are increasingly popular and deserve to be studied well (von Davier & Haberman, 2014). Yet, this is likely to be more involved since such models are not monotone without specific parameter constraints (van der Linden, 2012).

Acknowledgements

The authors would like to thank Shelby Haberman, Bob Mislevy, and the editor for helpful comments on earlier versions of the paper.

The research reported in this article was supported in part by the Institute of Education Sciences, U.S. Department of Education, through grant R305D110027 to Educational

Testing Service. The opinions expressed are those of the authors and do not represent the views of the Institute or the Department of Education.

References

Adams, R., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23. doi:10.1177/0146621697211001

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Vol. *V*, pp. 111–150). Berkeley: University of California Press.

Bartholomew, D. J. (1981). Posterior analysis of the factor model. *British Journal of Mathematical and Statistical Psychology*, *34*, 93–99. doi:10.1111/j.2044-8317.1981.tb00620.x

Bartholomew, D. J. (1984). The foundations of factor analysis. *Biometrika*, *71*, 221–232. doi:10.1093/biomet/71.2.221

Berkson, J. (1946). Limitations of the application of fourfold tables to hospital data. *Biometrics Bulletin*, *2*, 47–53. doi:10.2307/3002000

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

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

Bollen, K., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators and covariates. *Psychological Methods*, *16*, 265–284. doi:10.1037/a0024448

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory analysis: A study on the 5-factor model of personality. *Personality and Individual Differences*, *11*, 515–524. doi:10.1016/0191-8869(90)90065-Y

Darroch, J., Lauritzen, S., & Speed, T. (1980). Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics*, *8*, 522–539.

Díez, F., & Druzdzel, M. (2006). *Canonical probabilistic models for knowledge engineering*(Tech. Rep. No. CISIAD-06-01). Madrid: UNED.

Gibbons, R., & Hedeker, D. (1992). Full-information item bi-factor analyses. *Psychometrika*, *57*, 423–436. doi:10.1007/BF02295430

Gibbs, J. (1902). *Elementary principles in statistical mechanics*. New Haven, CT: Yale University Press.

Haberman, S. (2013). *A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm* (Research Report pp. 13–32). Princeton, NJ: Educational Testing Service.

Hedeker, D. (2005). Generalized linear mixed models. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 729–738). Chichester: Wiley.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210. doi:10.1007/s11336-008-9089-5

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

Holzinger, K., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*, 41–54. doi:10.1007/BF02287965

Hooker, G. (2010). On separable tests, correlated priors, and paradoxical results in multidimensional item response theory. *Psychometrika*, *75*, 694–707. doi:10.1007/S11336-010-9181-5

Hooker, G., & Finkelman, M. (2010). Paradoxical results and item bundles. *Psychometrika*, *75*, 249–271. doi:10.1007/S11336-009-9143-Y

Hooker, G., Finkelman, M., & Schwartzman, A. (2009). Paradoxical results in multidimensional item response theory. *Psychometrika*, *74*, 419–442. doi:10.1007/S11336-009-9111-6

Hoshino, T., & Shigemasu, K. (2008). Standard errors of estimated latent variable scores with estimated structural parameters. *Applied Psychological Measurement*, *32*, 181–189. doi:10.1177/0146621607301652

Jaakkola, T. S., & Jordan, M. I. (1999). Variational probabilistic inference and the QMRDT network. *Journal of Artificial Intelligence Research*, *10*, 291–322. doi:10.1613/jair.583

Joe, H. (1997). *Multivariate models and dependence concepts*. London: Chapman & Hall.

Jordan, P., & Spiess, M. (2012). Generalizations of paradoxical results in multidimensional item response theory. *Psychometrika*, *77*, 127–152. doi:10.1007/S11336-011-9243-3

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202. doi:10.1007/BF02289343

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272. doi:10.1177/01466210122032064

Karlin, S. (1968). *Total positivity* (Vol. 1). Stanford, CA: Stanford University Press.

Karlin, S., & Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities, I. Multivariate totally positive distributions. *Journal of Multivariate Analysis*, *10*, 467–498. doi:10.1016/0047-259X(80)90065-2

Krijnen, W. P., Wansbeek, T., & ten Berge, J. M. (1996). Best linear predictors for factor scores. *Communications is statistics: Theory and Methods*, *25*, 3013–3025. doi:10.1080/03610929608831883

Lauritzen, S. (1996). *Graphical models*. New York: Oxford University Press.

Lawley, D., & Maxwell, A. (1971). *Factor analysis as a statistical method*. London: Butterworths.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212. doi:10.1007/BF02294535

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

McDonald, R. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*, 99–114. doi:10.1177/01466210022031552

Mellenbergh, G. (2011). *A conceptual introduction to psychometrics*. The Hague: Eleven International.

Milgrom, P. (1981). Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, *12*, 380–391. doi:10.2307/3003562

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, *80*, 993–997. doi:10.1080/01621459.1985.10478215

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, *59*, 439–483. doi:10.1007/BF02294388

Neudecker, H. (2004). On best affine unbiased covariance-preserving prediction of factor scores. *SORT*, *28*, 27–36.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference* (2nd ed.). San Mateo, CA: Morgan Kaufmann.

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

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reckase, M. D., & Luo, X. (in press). A paradox by another name is good estimation. In A. van der Ark, D. Bolt, S.-M. Chow, J. Douglas & W.-C. Wang (Eds.), *Proceedings of IMPS 2014*. New York, NY: Springer.

Reise, S. (2012). The resdiscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696. doi:10.1080/00273171.2012.715555

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. doi:10.1111/j.1745-3984.2010.00118.x

Rijmen, F., & de Boeck, P. (2005). A relation between a between-item multidimensional IRT model and the mixture Rasch model. *Psychometrika*, *70*, 481–496. doi:10.1007/s11336-002-1007-7

Rijmen, F., Tuerlinckx, F., Boeck, P. D., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185–205. doi:10.1037/1082-989X.8.2.185

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, *53*, 349–359. doi:10.1007/BF02294217

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48* (2).

Schmid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61. doi:10.1007/BF02289209

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: CRC Press.

Smyth, P., Heckerman, D., & Jordan, M. (1997). Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, *9*, 227–269. doi:10.1162/neco.1997.9.2.227

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese, 48,* 191–199. doi:10.1007/BF01063886

Thurstone, L. (1947). *Multiple factor analysis: A development and expansion of the vectors of the mind*. Chicago: University of Chicago Press.

van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308. doi:10.1007/s11336-006-1478-z

van der Linden, W. (2012). On compensation in multidimensional response modeling. *Psychometrika*, *77*, 21–30. doi:10.1007/S11336-011-9237-1

van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika*, *67*, 519–538. doi:10.1007/BF02295129

van Rijn, P. W., Deane, P., Rijmen, F., & Bennett, R. E. (2014). *Considerations in fitting confirmatory MIRT models: An application to innovative writing assessment*. (Submitted for publication.)

van Rijn, P. W., & Rijmen, F. (2012). *A note on explaining away and paradoxical results in multidimensional item response theory* (ETS No. RR-12-13). Princeton, NJ: Educational Testing Service.

von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classifcation models morphing into unidimensional diagnostic classification models –A commentary. *Psychometrika*, *79*, 340–346. doi:10.1007/s11336-013-9363-z

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2007). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26 Psychometrics* (pp. 1039–1055). Amsterdam: Elsevier.

Wainer, H., Bradlow, E., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Wang, C. (2014). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika*. doi:10.1007/S11336-013-9399-0

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. doi:10.1007/BF02294627

Wellman, M., & Henrion, M. (1993). Explaining 'explaining away'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*, 287–292. doi:10.1109/34.204911

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, *60*, 181–198. doi:10.1007/BF02301412

Wright, S. (1921). Systems of mating. *Genetics*, *6*, 111–178.

Yung, Y.-F., Thissen, D., & McLeod, L. (1999). On the relationship between the higher-order model and the hierarchical factor model. *Psychometrika*, *64*, 113–128. doi:10.1007/BF02294531

Received 23 August 2013; revised version received 24 October 2014