Cognitive Diagnostic Assessment via Bayesian Evaluation of Informative Diagnostic Hypotheses

Herbert Hoijtink
CITO Institute for Educational Measurement and Department of Methods and Statistics, Utrecht University

Sébastien Béland
University of Sherbrooke

Jorine A. Vermeulen
University of Twente and CITO Institute for Educational Measurement

Author Note

CITO Institute for Educational Measurement, P.O.Box 1034, 6801 MG Arnhem, The Netherlands. Department of Methods and Statistics, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands. E-mail: herbert.hoijtink@cito.nl; h.hoijtink@uu.nl; sebastien.beland@usherbrooke.ca; j.a.vermeulen@utwente.nl. This research was supported by a grant from the Netherlands Organization for Scientific Research: NWO-VICI-453-05-002.

Abstract

There exist diverse approaches that can be used for cognitive diagnostic assessment, like mastery testing, constrained latent class analysis, rule space methodology, diagnostic cognitive modeling, and person fit analysis. Each of these approaches can be used within one of the four psychometric perspectives on diagnostic testing discussed by Borsboom (2008), that is, the dimensional, diagnostic, constructivist, and causal system perspectives. Bayesian evaluation of informative diagnostic hypotheses is an alternative for each of the other approaches that is more flexible in the diagnostic hypotheses that can be evaluated, and can be used in each of the four psychometric perspectives on diagnostic testing. After being formulated, informative diagnostic hypotheses are evaluated by means of the Bayes factor using only the data from the person to be diagnosed. Already relatively small diagnostic tests render Bayes factors that provide convincing evidence in favor of one of the diagnostic hypotheses under consideration.

Keywords: Bayes Factor, Cognitive Diagnostic Assessment, Inequality Constrained Hypothesis, Informative Hypothesis, Psychometrics

Cognitive Diagnostic Assessment via Bayesian Evaluation of Informative Diagnostic Hypotheses

Introduction

Cognitive diagnostic assessment is commonly used to diagnose or identify psychological states, levels of achievement/mastery, or cognitive deficits. This process is carried out via the evaluation of the responses of a person to items that are expected to reflect to which state, level, or deficit a person belongs. Cognitive diagnostic assessment occurs in a variety of situations. In this paper three examples will be used. The first example is placed in the context of educational measurement. The items are arithmetic exercises, and the diagnosis to be made is whether or not a person responds in accordance with a model based on the principle that the probability of a correct item response depends on the ability of the person and the difficult of the item. The second example is placed in the context of developmental psychology. Using balance scale tasks from Piaget (Boom, Hoijtink, & Kunnen, 2001; Rijkes & Kelderman, 2006; Siegler, 1981), the goal is to determine the developmental stage to which a child belongs. The third example is placed in the context of psychiatric evaluations. Using items from the Diagnostic and Statistical Manual of Mental Disorders IV-TR (DSM-IV-TR, American Psychiatric Association, 2000), the goal is to determine whether a person suffers from an avoidant personality disorder or not.

There is a wide body of literature about cognitive diagnostic assessment. Among these are mastery testing (Dayton & MacReady, 1988; MacReady & Dayton, 1977, 1980; Glas & Vos, 2010; Vos & Glas, 2010), rule space methodology (Tatsuoka, 1983; Tatsuoka & Tatsuoka, 1997; Tatsuoka, 2009), knowledge spaces (Doignon & Falmagne, 1999; Schrepp, 2005), constrained latent class analysis (Hoijtink, 2001), Q-matrix based classification (Chiu, Douglas, & Li, 2009; De Carlo, 2012; De La Torre, 2011; Liu, Xu, & Ying, 2012; Rupp, Templin, & Henson, 2010; Von Davier, 2011; Wang, Chang, & Douglas, 2012), and person fit analysis (Karabatsos, 2003; Meijer & Sijtsma, 2001). The interested reader is

referred to DiBello, Roussos, and Stout (2007), Leighton and Gierl (2007), and Rupp, Templin, and Henson (2010), for rather encompassing reviews and introductions.

This paper has three goals. First of all, it will be shown that each of the existing approaches is tailored to a specific form of cognitive diagnostic assessment. More specifically, each approach can be categorized into one of the four psychometric perspectives on cognitive diagnostic assessment as presented by Borsboom (2008): the dimensional perspective, the diagnostic perspective, the constructivist perspective, and the causal systems perspective. Secondly, a new approach called Bayesian evaluation of informative diagnostic hypotheses (BED) will be introduced. Where traditional null-hypotheses state that "nothing is going on" (e.g., all the means in a one-way analysis of variance are equal) and alternative hypotheses state that "something is going on but it is unclear what is going on" (e.g., not all the means in a one-way analysis of variance are equal), informative hypotheses state "what is going on" (e.g., the first mean is smaller than the second mean which in turn is smaller than the third mean etc.). As will be illustrated throughout this paper, informative *diagnostic* hypotheses can be used to describe the states, levels, and deficits of interest in cognitive diagnostic assessment. Thirdly it will be shown that Bayesian evaluation of informative diagnostic hypotheses provides an alternative for other approaches that can be used in each perspective, and provides added flexibility with respect to the diagnostic hypotheses that can be evaluated.

The paper is structured as follows. First, it is elaborated how informative diagnostic hypotheses can be constructed. Secondly, Bayesian evaluation of informative diagnostic hypotheses is introduced. Thereafter, the importance of predictive validity in the context of the evaluation of informative diagnostic hypotheses is highlighted. Subsequently, an example of BED is given for each psychometric perspective, including a comparison with other approaches that can be used in the perspective at hand. The paper continues with an evaluation of the performance of BED, is concluded with a short discussion, and contains an appendix in which software with which BED can be applied is described.

Informative Diagnostic Hypotheses

As is illustrated in Table 1, cognitive diagnostic assessment has a double context. On the one hand there is a domain within which persons have to be assessed. In this paper examples will be given concerning the measurement of arithmetic ability, the determination of developmental stages, and the diagnosis of an avoidant personality disorder. On the other hand, assessment takes place within one of four psychometric perspectives as elaborated by Borsboom (2008). Short characterizations (later in this paper more elaborate descriptions will be given) of each perspective are:

- The dimensional perspective: A diagnosis is based on a person's location on one or more latent continua.
- The diagnostic perspective: A diagnosis is based on a person's membership of one of a set of categorical latent classes.
- The constructivist perspective: A diagnosis is based on the responses of a person to a convenient grouping of items. In the example that will be introduced later on, a group of symptoms representing an avoidant personality disorder will be used.
- The causal system perspective: A diagnosis is based on the responses of a person to a causal grouping of items, that is, the response to an item is a function of the responses given to items preceding the current item in the causal system.

The construction of informative diagnostic hypotheses starts with the choice of a psychometric perspective. As is indicated in Table 1 usually the perspective follows naturally from the domain of interest. To give two examples: Arithmetic ability is often represented as a latent continuum, which leads to the dimensional perspective; and developmental stages are often represented as latent classes, which leads to the diagnostic perspective. As a result within different domains different traditions have originated. The dimensional perspective is often used in educational testing (Jacob & Levitt, 2003; Jaeger, 1988; Meijer, Egberink, Emons, & Sijtsma, 2008). The diagnostic perspective is used to determine the developmental stage a person belongs to. An example will be given using

Piaget's balance scale task, more examples can be found at http://en.wikipedia.org/wiki/Child_development_stages, where stages of development for various skills are described. The constructivist perspective is omnipresent in psychological testing using the DSM-IV-TR (American Psychiatric Association, 2000). The causal systems perspective (Borsboom, 2008) is rather new, and has sofar been used by Borsboom, Epskamp, Kievit, Cramer and Schmittmann (2011), Cramer, Waldorp, van der Maas, Borsboom (2010), and Schmittmann, Cramer, Waldorp, Epskamp, Kievit, Borsboom (2011). Especially in the context of psychological and psychiatric evaluations it provides an alternative for the constructivist perspective.

After the choice of a psychometric perspective, informative diagnostic hypotheses have to be constructed. In this section and in the next two sections important steps in the construction process will be discussed and illustrated using a small arithmetic ability test evaluated in the dimensional perspective. Examples for the other perspectives will be presented later in this paper. Consider the following arithmetic ability test:

1. $1 + 4 =$
2. $8 + 21 =$
3. $33 + 56 =$
4. $443 + 43 =$
5. $743 + 533 =$
6. $243 + 4354 =$,

where each item response x_j for $j = 1, ..., J$ (in this example $J = 6$) is scored 0 (incorrect) or 1 (correct).

First of all, one or more cognitive models have to be formulated, that is, the assumptions with respect to the response process have to be stated. For the arithmetic ability test there is one simple cognitive model: Each item has a difficulty depending on the size of the numbers that have to be added; each pupil has an ability; and, the probability of a correct response π_j for $j = 1, ..., 6$ depends on the difficulty of an item (the larger the size

of the numbers that have to be added the smaller the probability of a correct response) and the ability of the pupil (the higher the location of the person on a latent ability continuum, the higher the probability of a correct response). This means that the π_j's are pupil specific parameters. However to avoid unnecessary heavy notation the "pupil dependence" of the π_j's is left implicit. This cognitive model is the basis of many item response models (Birnbaum, 1968; De Boeck & Wilson, 2004; Embretson & Reise, 2000; Fischer & Molenaar, 1995; Fox, 2010; Hambleton & Swaminathan, 1985; Lord, 1980; Sijtsma & Molenaar, 2002). Each π_j has a frequency interpretation: If many persons with the same ability respond to item j, a proportion π_j of these persons will give the correct response.

In the arithmetic ability example ability estimates (like the number of correct responses) are only adequate if a pupil responds according to the assumptions specified in the cognitive model. Not responding in agreement with the cognitive model may, for example, be caused by cheating (copying answers to more difficult items from a neighbor) or loss of concentration (not being challenged enough by the easier items). The diagnostic question of interest is therefore whether or not a pupil responds in agreement with the cognitive model. In general, the diagnostic question of interest is which of the cognitive models under consideration provides the best explanation of the responses that a person has given. In the example concerning developmental stages it will be illustrated that there may be more than one cognitive model.

Secondly, each cognitive model has to be translated into an informative diagnostic hypothesis. In this and the following sections informative diagnostic hypotheses are introduced by means of examples. Later in this paper we will present the general form of the informative diagnostic hypotheses that can be handled with the approach proposed. The basic idea is to specify hypotheses using inequality constraints ($<$ denoting smaller than and $>$ denoting larger than) among the response probabilities π_j. Although the π_j's can not be estimated for one person, it is, as will be shown, possible to evaluate whether or not the π_j's for a person are in agreement with an informative diagnostic hypothesis.

Before hypotheses can be specified expert judgement possibly in combination with calibration is needed to ensure that the main components of each cognitive model are well-defined. Here this will be elaborated for the arithmetic ability example within the dimensional perspective, later in this paper this will be elaborated for the other psychometric perspectives. A key feature of the cognitive model for the arithmetic ability example is that the items have to be ordered according to difficulty. This can be done in two manners: expert evaluation or calibration. An expert might argue that the difficulty of an item depends on the number of digits used in the numbers that have to be added. This would imply that Item 1 is the easiest and Item 6 the most difficult item. An advantage of this approach is that only the responses of the pupil to be evaluated are needed in order to be able to obtain an evaluation of the pupil. This is convenient if a calibration sample is hard or expensive to obtain. A potential disadvantage of this approach is that the expert or experts may misjudge the difficulty of the items. In that case the cognitive model is to a lesser or greater degree misspecified. If the test is given to 6 year old boys, the responses of a calibration sample of 6 year old boys can be used to estimate the difficulty of each item (for example the proportion of correct responses given to the item) and to order the items according to the proportion of correct responses. The challenge for this approach is to define the contours (6 year old boys) of the calibration sample such that it corresponds to the group of pupils that have to be evaluated.

With an item order available the cognitive model is sufficiently specified to be able to formulate an informative diagnostic hypothesis. This is not an exact science. Using the expert judgements, the hypothesis that a pupil responds in accordance with the cognitive model can be specified as

$$H_{cog} : \pi_1 > \pi_2 > \pi_3 > \pi_4 > \pi_5 > \pi_6, \tag{1}$$

and the hypothesis that a pupil does not respond according to the cognitive model as

$$H_{not\ cog} : \text{not } H_{cog}. \tag{2}$$

An alternative formulation of H_{cog} could be

$$H_{cog}: \begin{array}{c} \pi_1 > \{\pi_3, \ldots, \pi_6\} \\ \pi_2 > \{\pi_4, \pi_5, \pi_6\} \\ \pi_3 > \{\pi_5, \pi_6\} \\ \pi_4 > \{\pi_6\} \end{array}. \tag{3}$$

This hypothesis accounts for the fact that items adjacent in the ordering may have similar item difficulties which would imply that $\pi_j > \pi_{j+1}$ for $j = 1, ..., 5$ is a rather strict requirement. Still other formalizations of the hypothesis that a pupil responds in accordance with the cognitive model are conceivable. Which formalization of H_{cog} is the best, depends on its predictive validity. After the next section it will be elaborated how an experiment can be used to compute the proportion of correct classifications of pupils that do and do not respond in accordance with the cognitive model. If an experiment is not feasible, the best that can be done is to achieve agreement among the peers working in a certain domain about the best manner to formalize an informative diagnostic hypothesis.

After choosing a psychometric perspective and formulating and formalizing cognitive models, the models have to be evaluated for each of the pupils under investigation. The Bayesian methodology with which this can be done will be introduced in the next section: It will be shown how for each pupil the most likely informative diagnostic hypothesis can be determined, and the statistical meaning of the phrase "the most likely informative diagnostic hypothesis" will be elaborated. This section will end by giving a methodological context for and interpretation of this phrase.

Within each cognitive domain of interest there are potentially many hypotheses that can be evaluated. However, only a small subset of all possible hypotheses is actually evaluated, which is a substantial reduction of the problem space. This is good because these hypotheses reflect the key decisions that have to be taken with respect to the persons that are to be diagnosed. Stated otherwise, these hypotheses represent the key cognitive strategies in which the assessor is interested. However, it is important to realize that the

phrase “the most likely informative hypothesis” consequently refers to the most likely of the informative hypotheses under investigation. This has a number of implications. There may be other cognitive models or response strategies that are not considered by or unknown to the assessor that would give better explanations of the item responses than the cognitive models represented by the informative hypotheses under consideration. This means that the phrase “the most likely informative hypothesis” assumes that the test taking behavior of the persons to be assessed is as intended, that is, is in accordance with one of the cognitive models specified by the assessor. A related issue is addressed in the work of Kripke (1982). No number of item responses can uniquely identify an informative hypothesis. There can always be another hypothesis that gives an equally good or better explanation of the item responses. This highlights another assumption of the approach proposed. Paraphrasing George Box’s quote “all models are wrong but some are useful” it can be said that all diagnostic hypotheses are assumed to be wrong but may nevertheless be useful because they are close enough to the true hypothesis to enable assessors to reach key decisions with respect to the persons to be assessed. Finally, it has to be realized that all informative diagnostic hypotheses under consideration can be wrong, that is, even the best of a set of informative diagnostic hypotheses may be not useful.

There are two manners in which the last issue can be addressed. First of all, an informative diagnostic hypotheses can be compared with its complement as is done in Equations 1 and 2. In this way all possible combinations of the π_j’s are covered, that is, the “best” combination is through the manner in which the hypotheses are formulated always included by one of the hypotheses. Secondly, as will be illustrated in the context of the diagnostic perspective, each informative diagnostic hypothesis can be compared to an unconstrained hypothesis H_u. If in terms of the Bayes factor (to be elaborated in the next section) an informative diagnostic hypothesis performs worse than an unconstrained hypothesis, it can be concluded that the constraints are not supported by the data, that is, that the hypothesis is not a useful (in terms of Box’s quote) explanation of the item

responses.

Bayesian Evaluation of Informative Diagnostic Hypotheses

The Bayes factor (Kass & Raftery, 1995; Lavine & Schervish, 1999) is a Bayesian criterion that can be used to select the best of a set of competing inequality constrained hypotheses (Hoijtink, 2012) via an evaluation of the complexity and fit of the hypotheses under investigation. In this paper informative diagnostic hypotheses are evaluated for each person, that is, Bayes factors are computed for each person. The procedure is the same for each person irrespective of whether one or many persons have to be evaluated. In the next three subsections complexity, fit, and the Bayes factor will shortly be introduced and illustrated. The interested reader is referred to Chapters 3, 4, and 10 from Hoijtink (2012) for a complete elaboration, motivation, and illustration of the Bayes factor as a tool for the evaluation of inequality constrained hypotheses.

Complexity

The prior distribution (Gelman, Carlin, Stern, & Rubin, 2004; Lynch, 2007) for the π_j's is the product of J independent, identical, and non-informative uniform distributions on the interval [0–1]:

$$h(\boldsymbol{\pi}) = \prod_{j=1}^{J} \text{Beta}(\pi_j \mid 1, 1) = 1, \tag{4}$$

where $\boldsymbol{\pi} = [\pi_1, ..., \pi_J]$. In the literature with respect to Bayesian evaluation of inequality constrained hypotheses (Hoijtink, 2012; Klugkist, Laudy, & Hoijtink, 2005; Mulder, Hoijtink, & Klugkist, 2010), the proportion of the prior distribution c_m in agreement with H_m, that is,

$$c_m = \int_{\boldsymbol{\pi} \in H_m} h(\boldsymbol{\pi}) \partial \boldsymbol{\pi}, \tag{5}$$

is called the complexity of the constrained hypothesis for $m = 1, ..., M$, where M denotes the number of informative diagnostic hypotheses. Note that c_m is estimated using a sample

of $\boldsymbol{\pi}$ from the prior distribution. The proportion of sampled $\boldsymbol{\pi}$'s in agreement with H_m is the estimate of c_m. Consider the following example:

Example 1. Let $H_{cog} : \pi_1 > \ldots > \pi_J$ and $H_{not\ cog}$: not H_{cog}. As may be clear, H_{cog} is more specific (it allows only a single ordering of the π_j's) than $H_{not\ cog}$ (which allows $J! - 1$ orderings of the π_j's). This is adequately reflected by complexities $c_{cog} = 1/J!$ and $c_{not\ cog} = (J! - 1)/J!$ both computed using Equation 5.

The prior distribution displayed in Equation 4 has two important properties. First of all, it is noninformative, in the sense that the posterior distribution (see the next subsection) that will be constructed using Equation 4 is proportional to the density of the data upon which it is based. Stated otherwise, the posterior distribution is determined by the data, and in this sense objective. Secondly, it is *not* used to formalize prior knowledge, but it is used to quantify the complexity of the inequality constrained hypotheses under investigation.

Fit

The posterior distribution (Gelman, Carlin, Stern, & Rubin, 2004; Lynch 2007) of the π_j's is proportional to the product of the density of the data based on local independence of the item responses

$$f(\boldsymbol{x} \mid \boldsymbol{\pi}) = \prod_{j=1}^{J} \pi_j^{x_j} (1 - \pi_j)^{1 - x_j}, \tag{6}$$

where $\boldsymbol{x} = [x_1, ..., x_J]$, and $x_j \in \{0, 1\}$ for $j = 1, ...J$, and the prior distribution displayed in Equation 4:

$$g(\boldsymbol{\pi} \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \boldsymbol{\pi}) h(\boldsymbol{\pi}) \propto$$

$$\prod_{j=1}^{J} \text{Beta}(\pi_j \mid 1 + x_j, 1 + (1 - x_j)). \tag{7}$$

The proportion of the posterior distribution in agreement with H_m, that is,

$$f_m = \int_{\boldsymbol{\pi} \in H_m} g(\boldsymbol{\pi} \mid \boldsymbol{x}) \partial \boldsymbol{\pi}, \tag{8}$$

is called the fit of the constrained hypothesis at hand (Klugkist, Laudy, & Hoijtink, 2005; Mulder, Hoijtink, & Klugkist, 2010). Note that f_m is estimated using a sample of $\boldsymbol{\pi}$ from the posterior distribution. The proportion of sampled $\boldsymbol{\pi}$'s in agreement with H_m is the estimate of f_m. Consider the following example:

Example 2. Let $H_{cog} : \pi_1 > \ldots > \pi_6$, let $\boldsymbol{x}_1 = [111000]$, and let $\boldsymbol{x}_2 = [000111]$. As can be seen, $\boldsymbol{x}_1$ is in agreement with H_{cog}, that is, easy items are responded to correctly, and more difficult items are responded to incorrectly, while $\boldsymbol{x}_2$ is not. This is nicely reflected by the fit of H_{cog} computed using Equation 8 for both response vectors: $f_{cog|\boldsymbol{x}_1} = .01102$ and $f_{cog|\boldsymbol{x}_2} = .00004$.

The Bayes Factor

The Bayes factor (Kass & Raftery, 1995; Lavine & Schervish, 1999) is a quantification of the relative support in the data for two competing hypotheses. For example, Klugkist, Laudy, and Hoijtink (2005), Mulder, Hoijtink, and Klugkist (2010), and Hoijtink (2012), have extensively discussed the (derivation of) Bayes factors for the evaluation of inequality constrained hypotheses like displayed in Equations 1, 2, and 3.

Three forms of the Bayes factor will be used in this paper. The first form compares an informative diagnostic hypothesis H_m with a hypothesis without constraints on the π_j's which is denoted by $H_u : \pi_1, ..., \pi_J$ where the subscript u denotes that the π_j's are unconstrained:

$$BF_{mu} = \frac{f_m}{c_m}. \tag{9}$$

The Bayes factor can also be used to evaluate two competing informative diagnostic hypotheses H_m and $H_{m'}$ with respect to the process that generated a person's responses:

$$BF_{mm'} = BF_{mu}/BF_{m'u} = \frac{f_m}{c_m}/\frac{f_{m'}}{c_{m'}}. \tag{10}$$

If $H_{m'} = H_c$, that is, the complement of H_m, the Bayes factor can be written as

$$BF_{mc} = \frac{f_m}{c_m}/\frac{1 - f_m}{1 - c_m}. \tag{11}$$

In the Appendix to this paper the program BED.exe is presented. It can be downloaded from http://tinyurl.com/hoijtinkbook under the rubric "New Developments and Software". Based on a text input file containing the hypothesis to be evaluated and the response vector of the person to be assessed, the program computes BF_{mu} and BF_{mc}. Since there is uncertainty in the estimates of c_m and f_m due to sampling, there is also uncertainty in the estimates of BF_{mu} and BF_{mc}. Using the approach presented in Chapter 10 of Hoijtink (2012), BED.exe computes a 95% Monte Carlo interval representing this uncertainty. If the interval is considered to be too large, the size of the samples obtained from the prior and posterior distribution can be increased to obtain a smaller Monte Carlo interval.

Example 3. If $\boldsymbol{x} = [111000]$, $H_{cog} : \pi_1 > \ldots > \pi_6$, and $H_{not\ cog}$: not H_{cog}, then $c_{cog} = 1/6!$, $c_{not\ cog} = 1 - 1/6!$, $f_{cog} = .011$, and $f_{not\ cog} = 1 - .011$. This results in a Bayes factor $BF_{cog,not\ cog} = 8.02$, which shows that there is 8.02 times more support for H_{cog} than for $H_{not\ cog}$.

The number 1 is a natural reference value. $BF_{mm'} > 1$ represents more support for H_m, and $BF_{mm'} < 1$ represents more support for $H_{m'}$. The degree of support is determined by the size of the Bayes factor. For example, $BF_{mm'} = 5$ (the support for m is 5 times stronger than the support for m') reflects more support for m compared to m' than $BF_{mm'} = 1.2$. Similarly, $BF_{mm'} = .1$ reflects more support for m' compared to m than $BF_{mm'} = .9$. Kass and Raftery (1995) provide guidelines in the line of Jeffreys (1961) for the interpretation of the Bayes factor. Values of 1–3 (or 1/3–1) are "not worth more than a bare mentioning", values of 3–20 (or 1/20–1/3) are "positive evidence", values of 20–150 (or 1/150–1/20) are "strong evidence", and values over 150 (or under 1/150) are "very strong evidence". Bayes factors can be used to infer from the item responses given by the person which hypothesis gives the best description of the person. Using these guidelines, it can in Example 3 be concluded that there is positive evidence in favor of H_{cog}, that is, based on

the item responses it is inferred that H_{cog} is the best of the hypotheses under consideration. As elaborated earlier, the hypotheses under consideration represent the key decisions that an assessor has to make. This implies that there could be another hypothesis that provides an even better explanation of the item responses. The approach proposed does not render "the objective truth" based on the responses to a set of items. However, it does render the best of a set of competing hypotheses that are meaningful in a specific assessment context.

Predictive Validity

In the previous sections it has been described and illustrated how a diagnostic evaluation of a person can be obtained. An important issue is the predictive validity or predictive utility of this diagnosis for the population of interest, that is, to what extend does the evaluation of persons using item responses and informative diagnostic hypotheses render the correct diagnosis. Note that a test based diagnostic evaluation having a low predictive validity does not imply that the test is also deficient in other kinds of validity like internal, content, and construct validity (Carmines & Zeller, 1979). However, it does mean that use of the test does not contribute to a diagnostic evaluation of the persons in the population of interest.

Based on Meehl and Rosen (1955), Schonemann and Thompson (1996), Schonemann (1997; 2005), and Taylor and Russell (1939) (visit http://www.schonemann.de/publications.htm for easy access to the Schonemann publications) it will now be discussed how predictive validity can be assessed. This includes a discussion of the use of predictive validity as an argument in the choice between different formalizations of the same cognitive model as in Equations 1 and 3.

Predictive validity can be determined via the execution of the following four steps:

1. Use expert evaluations to create groups of persons with diagnosis m formalized as H_m for $m = 1, \ldots, M$. Ascertain that these groups of persons correspond to the population of interest with respect to variables such as gender and age.

2. Collect the responses of each person to the J items that constitute the diagnostic test.

3. Use diagnostic testing (in this paper based on the evaluation of informative diagnostic hypotheses using the Bayes factor) to evaluate the item responses of each person, and assign each person to one of the M hypotheses.

4. Construct a contingency table with in the rows the true diagnosis (obtained in the first step) and in the columns the predicted diagnosis (obtained in the third step) of each person. Table 2 presents an example for two hypotheses m and m'. Adjusting the labeling of Schonemann and Thompson (1996) to the context of this paper the entries of this table can be labeled T_m (true classifications to H_m), F_m (false classifications to H_m), $T_{m'}$ (true classifications to $H_{m'}$), and $F_{m'}$ (false classifications to $H_{m'}$).

There are three pitfalls that should be avoided when interpreting this table. First of all, as elaborated in Taylor and Russell (1939), Schonemann and Thompson (1996), and Schonemann (1997), the validity coefficient (e.g., the Pearson correlation computed using the entries of Table 2) is often a poor indicator of the effectiveness of diagnostic tests. This is due to the fact that a correlation is a rather indirect manner to quantify the number of correct classifications as presented in Table 2. Another popular quantity is the proportion of correct classifications PCC $= \frac{T_m+T_{m'}}{T_m+T_{m'}+F_m+F_{m'}}$. However, as elaborated by Meehl and Rosen (1955), Schonemann and Thompson (1996), and Schonemann (1997, 2005), this quantity may lead to a second pitfall: the base rate confound. Suppose, for example, that $T_m = 1$, $F_m = 9$, $T_{m'} = 90$, and $F_{m'} = 10$. This leads to PCC $= 91/110 = .83$. Although this seems to be a reasonable number, it is not. Imagine that (without using a test) all the persons would be classified in $H_{m'}$. This would render PCC $= 99/110 = .90$. Stated otherwise, not using a test but simply assigning all the persons to the hypothesis with the highest base rate ($H_{m'}$ is more frequent than H_m) gives better results than our test based classification. The conclusion must be that the predictive validity of the test is insufficient.

A better approach is to use two conditional probabilities to evaluate tables like Table

2: $\text{PC}_m = \frac{T_m}{T_m+F_{m'}}$ and $\text{PC}_{m'} = \frac{T_{m'}}{T_{m'}+F_m}$, that is, the probabilities of being correctly diagnosed if the true hypotheses are H_m and $H_{m'}$, respectively. For Table 2 $\text{PC}_m = 1/11 = .09$ and $\text{PC}_{m'} = 90/99 = .91$. It is now immediately clear that the predictive validity of the test for persons belonging to H_m is very low.

A third pitfall is highlighted by Schonemann and Thompson (1996), and is called the hit-rate bias. A (diagnostic) test can be biased against specific subgroups. It could, for example, be that for the items presented in Table 3, which can be used to diagnose an avoidant personality disorder, the response "applies" is used more often for men than for women even if they show the same degree of avoidance. If this is true the predictive validity will not be the same for men and women. It is therefore recommended to compute and evaluate Table 2 not only for a sample from the population of interest as a whole, but also for the most relevant subgroups.

As was illustrated in the arithmetic ability example, there may be more than one way in which a cognitive model can be formalized into informative diagnostic hypotheses (see, for example, Equations 1 and 3. Which of the formalizations under consideration is the best, can be decided if their predictive validity is evaluated. The formalization with the smallest error rates is best able to distinguish the cognitive models under consideration using the item responses. If none of the formalizations under consideration have a sufficient predictive validity, assessors have to reevaluate their items and their informative diagnostic hypotheses.

Informative Diagnostic Hypotheses in Four Psychometric Perspectives

In the previous sections Bayesian evaluation of informative diagnostic hypotheses has been introduced. Using an arithmetic ability test, it was illustrated that first of all a psychometric perspective has to be chosen. Subsequently, cognitive models for the response behavior of the person to be evaluated have to be constructed and formalized into informative diagnostic hypotheses. Thereafter, the Bayes factor is used to evaluate the

hypotheses of interest. Finally, the predictive validity of diagnostic test and hypotheses has to be considered.

What has not yet been done is compare the approach proposed to other approaches that can be used within the dimensional perspective. This will be remedied in the next section. Thereafter, examples for the diagnostic, constructivist, and causal systems perspectives will be presented and discussed along the same lines as used for the arithmetic ability example. This section will be concluded with a subsection in which the general form of the informative diagnostic hypothesis that can be evaluated using the approach elaborated in this paper is presented and a subsection that will provide a short comparison of the statistical tools that are used in various approaches: the Bayes factor, p-values, and classification probabilities.

The Dimensional Perspective

Within the dimensional perspective the approach proposed in this paper resembles person fit analysis (PFIT) as is used in item response models. Invariably two ingredients are involved in PFIT: The response vector of a person e.g. 111100, that is, a person responding correctly to the first four and incorrectly to the last two arithmetic ability items; and, the vector $[\pi_1, \ldots, \pi_6]$ containing, for the person at hand, the unknown probability of a correct response for each of the items. A calibration sample is needed to be able to compute these probabilities. In, for example, the Rasch model, the responses of N persons to J items are used to estimate the ability of each person (denoted by θ for the person at hand) and the difficulty of each item δ_j for $j = 1, \ldots, J$. Using these estimates the required probabilities are computed by

$$\pi_j = \frac{exp(\theta - \delta_j)}{1 + exp(\theta - \delta_j)} \text{ for } j = 1, \ldots, J. \tag{12}$$

Subsequently the null-hypothesis H_0 : The person responds according to the Rasch model is evaluated against the alternative H_a : The person does not respond according to the Rasch model. This is often done by means of a p-value based on one of many available test

statistics. The interested reader is referred to Meijer and Sijtsma (2001) and Karabatsos (2003) for encompassing reviews.

The main differences between BED and PFIT are summarized in Table 4. Although a calibration sample may be useful in the context of BED, it is not necessary if there is a clear idea about the order of the items on the latent trait. Furthermore, the use of inequality constraints in the construction of informative hypotheses makes BED more flexible in the hypotheses that can be formulated and evaluated than PFIT. In PFIT the nul-hypothesis is invariably a counterpart of Equation 1 (see Klauer, 1991, 1995, for exceptions) whereas in BED variations like the hypothesis displayed in Equation 3 are possible. Another feature that further increases the flexibility with which hypotheses can be constructed is the use of logical operators. This feature will be introduced below in the subsection addressing the causal systems perspective.

Another difference between BED and PFIT is that BED uses the Bayes factor to evaluate hypotheses whereas PFIT uses p-values. The implication of this difference will further be elaborated in the last subsection of this section.

The Diagnostic Perspective

In the diagnostic perspective a diagnosis is based on a person's membership of one of a set of categorical latent classes (McCutcheon, 1987). Each class corresponds to a condition that underlies and explains the item responses given by the person at hand. These classes are unobservable, a person's class membership has to be inferred from the relation between the item responses given and the probabilities of responding correctly to each of the items in each latent class as specified by the latent class model at hand.

A well-known experiment from the psychological literature is the balance scale task of Piaget (Boom, Hoijtink, & Kunnen, 2001; Rijkes & Kelderman, 2006; Siegler, 1981). A picture of a simplified balance scale is shown to children. While the beam is fixed, a number of identical weights are placed on each side at certain distances from the fulcrum.

For each of a number of balances (the items) the children have to predict which side will tip, if any. The weights on the balance differ with respect to their number and distance to the center. The formal (torque) rule to obtain the correct answer is that the balance is in equilibrium when the product of the number of weights and the distance from the center is equal at both sides of the balance.

Here 8 items from the 19 balance scale items previously analyzed in Hoijtink and Boom (2008) and Laudy, Boom, and Hoijtink (2004) are used. These items are described in Table 5. Items 1 through 4 are "conflict weight" items. These are items for which both sides differ in distance and weight, and the correct answer is that the side with the most weight goes down. Items 5 through 8 are "conflict balance" items. For these items both sides also differ with respect to distance and weight. However, overall, the balance is in equilibrium.

According to Siegler's theory (Siegler, 1981), children use one of three strategies to respond to these items. Siegler's "Rule 1" states that children will look at the number of weights on each arm, and will predict that the arm with the largest number of weights will go down. This will lead to correct answers to items 1 through 4 and to incorrect answers to items 5 through 8. Siegler's "Rule 2" states that children look at both distance and weight. However, if both vary, they will give a random prediction. This will lead to random answers to each of the eight items presented in Table 5. If children use "Rule 3", that is, the torque rule, they will provide the correct answer to each of the items. Scoring the eight items in Table 5 for a hypothetical child rendered the responses 10110000.

The three rules of Siegler result from a carefully constructed experiment, and can straightforwardly be formalized in hypotheses:

$$H_{Rule1} : \pi_j > d \text{ for } j = 1, \ldots, 4, \text{ and } \pi_j < c \text{ for } j = 5, \ldots, 8, \tag{13}$$

$$H_{Rule2} : .33 - e < \pi_j < .33 + e \text{ for } j = 1, \ldots, 8, \tag{14}$$

$$H_{Rule3} : \pi_j > d \text{ for } j = 1, \ldots, 8. \tag{15}$$

The remaining question is how to choose c, d, and e. Both the elicitation of expert opinion and fine tuning by means of an evaluation of the predictive validity of informative diagnostic hypotheses constructed using different values for c, d, and e can be considered.

Experts may argue that, although they will relatively often give the correct response, it is unlikely that children who use the torque rule will always give the correct response. Stated otherwise, experts have to determine what the minimal probability of a correct response will be for a child that uses the torque rule. Values for d may be .75, .80, or even .90. In the same vein, it is also unlikely that children who use Rule 1 will always give the wrong answer to items 5 through 8. Sensible upper bounds on the probability of a correct response, that is values for c, may be .10, .20, and .25. It is also likely that in case of random responding the success probability will be about .33 (the alternatives children can choose from are *left side goes down*, *right side goes down*, and *equilibrium*). This can be reflected using values for e of .05 or .10.

Which of the set of reference probabilities is the best can be verified via determination of the predictive validity of the resulting set of hypotheses. If the time or money to evaluate predictive validity is lacking, one can also execute a sensitivity analysis to determine the effect of using different sets of reference values. An example will be given below. This topic will return when discussing Table 7 later in this paper.

For the response vector 10110000 the Bayes factors $BF_{Rule1,u}$ for hypotheses specified using c/d equal to .10/.90, .20/.80, and .25/.75 are 8.18, 11.58, and 12,17, respectively. As can be seen the conclusion is not sensitive to the choice of c and d. Using the guidelines presented by Kass and Raftery (1995) it can be concluded that Rule 1 is positively supported by the data. $BF_{Rule2,u}$ for e equal to .05 and .10 is 1.30 and 1.17, respectively, not sensitive to the choice of e: The data are not decisive with respect to Rule 2. $BF_{Rule3,u}$ for values of d equal to .90, .80, and .75, is .00, .001, and .005, respectively. Again the conclusion is not sensitive to the choice of d: Comparing Rule 3 with H_u there is very strong support for H_u. In summary, it may be concluded that the evaluation of Rule 1,

Rule 2, and Rule 3, is, in this example, not sensitive to the specification of reference values like c, d, and e, as long as the values chosen are reasonable given the rules for which they are used. As will become clear when Table 7 is discussed, the same holds for other response vectors.

Using $c = .20$, $d = .80$, and $e = .05$ the Bayes factor of H_{Rule1} versus H_{Rule2} equals 9.89. The Bayes factor of H_{Rule1} versus H_{Rule3} equals 11580.00. Using the guidelines from Kass and Raftery (1995) it can be concluded that the support in the data for H_{Rule_1} is strong to very strong. As elaborated earlier in this paper, these conclusions only hold if the test taking behavior of the child to be assessed is as intended, that is, if H_{Rule1}, H_{Rule2}, and H_{Rule3} cover the cognitive strategies used by a child. To avoid choosing the best of a set of useless hypotheses, each hypothesis can be compared with the unconstrained hypothesis H_u. This was done in the sensitivity analysis described above. It was clear that Rule 1 was positively preferred over H_u. This makes H_{Rule1} not only the best of the hypotheses under consideration but also a useful hypothesis in terms of Box's quote "All models are wrong but some are useful".

With respect to the latter, it is interesting to highlight the work by Rijkes and Kelderman (2006). Where our approach assumes that each child uses one of three strategies, their approach allows the strategy used to depend on the difficulty of the items and the ability of a child (the larger the ability the more often a more appropriate strategy will be chosen). Stated otherwise, if it is correct that a child uses a fixed strategy (which is the usual assumption in stagewise developmental research) our interpretation of the Bayes factor is fine. If, however, cognitive strategies are more along the lines of the model presented by Rijkes and Kelderman (2006), their model should be used, and each child can then be characterized by one ability parameter.

There are four other approaches that fit within the diagnostic perspective: constrained latent class analysis (CLCA; Hoijtink, 2001; Vermunt & Magidson, 2005); rule space methodology (RS; Tatsuoka, 1983; Tatsuoka & Tatsuoka, 1997; Tatsuoka, 2009);

knowledge structures (KS; Doignon & Falmagne, 1999; Schrepp, 2005); and diagnostic cognitive modelling (DCM; Chiu, Douglas, & Li, 2009; De Carlo, 2012; De La Torre, 2011; Leighton & Gierl, 2007; Liu, Xu, & Ying, 2012; Maris, 1999; Rupp, Templin, & Henson, 2010; Von Davier, 2011; Wang, Chang, & Douglas, 2012).

Within each approach three steps can be distinguished:

1. Determine the rules that define the set of categorical latent classes.
2. Use a calibration sample to fit a latent class model to the data.
3. Assign each person to one of the categorical latent classes.

CLCA is closely related to BED. In the example at hand, the rules that define the latent classes are given in Equations 13, 14, and 15. Subsequently, a calibration sample is used to fit a latent class model such that within each class the probabilities of responding correctly to an item are in agreement with these rules. Finally, for each person the probabilities of being classified in each class are computed.

RS and KS encompass various related approaches and elaborations. Here we will focus on KS as elaborated by Schrepp (2005). Rules are used to specify prototypical response vectors. A prototypical person responding according to $Rule1$ will render the response vector 11110000. For $Rule3$ 11111111 is the prototypical response vector. $Rule2$ states that a person responds randomly to each item with a probability of about .33 to give the correct response. This rule can not be captured in one prototypical response vector. If three correct out of eight responses is considered to be prototypical for this rule, there are eight-choose-three, that is, $c(8, 3) = 56$ response vectors for which this is true. Among them are 10110000 and 01001010. Subsequently, a constrained latent class model is formulated such that each class represents one of the prototypical response vectors. For example, 11110000 can be represented by a class in which $\pi_j > .9$ for $j = 1, \ldots, 4$ and $\pi_j < .1$ for $j = 5, \ldots, 8$. This model can be fitted using a calibration sample. Finally, for each person the probabilities of being classified in each class are computed. Note that this illustrates the flexibility of BED compared to RS and KS. With BED three rules are

sufficient to formalize the relevant cognitive models, with RS and KS $56 + 2 = 58$ prototypical response vectors are needed. Furthermore, it is questionable whether the set of 56 prototypical response vectors used to represent Rule 2 is adequate. Should, for example, also response vectors with four correct responses be included?

DCM uses a so-called Q-matrix to specify the relation between items and person attributes that are needed in order to be able to give the correct response to an item. In the example at hand two attributes can be distinguished: Is the item a conflict weight item or not, and is the item a conflict balance item or not. The corresponding Q-matrix is presented in the last two columns of Table 5. Note that DCM can be seen as a multivariate elaboration of mastery testing, that is, mastery of multiple instead of one attribute is evaluated. If more than one attribute influences the response to an item (not in the example) extra options become available. In, for example, deterministic-input-noisy-and (DINA) models all attributes related to an item have to be mastered in order to give the correct response. In deterministic-input-noisy-or (DINO) models at least one of the attributes has to be mastered in order to obtain the score 1 on an item. Note that Von Davier (2009) highlights why the linear logistic test model (LLTM; Fischer, 1973) is not a cognitive diagnostic model, despite the fact that, like the DCM, it also uses a Q-matrix. The main reason is that the LLTM relates only the items and not the persons to attributes.

Based on, for example, the DINA model, a latent class model can be formulated and fitted using a calibration sample. Within each class the probabilities of responding correctly to each item are estimated. The probabilities depend on the attributes relevant for an item, and which attributes are mastered by the persons in each class. The two attributes in Table 5 lead to four classes: Those who mastered both, one, or none of the attributes. This illustrates the relative flexibility of BED compared to DCM. For BED the relevant cognitive models can be captured in three informative diagnostic hypotheses, in standard DCM one is required to work with four latent classes. In the final stage of applying DCM for each person the probabilities of being classified in each class are computed.

Each of these approaches have and deserve a place in the literature with respect to cognitive diagnostic assessment. Many kinds of assessment can be executed using these approaches. However, as is summarized in Table 4, and illustrated in this subsection, BED is the more flexible approach. Comparing it to CLCA, it can be observed that it does not need a calibration sample, and it can use logical operators (introduced after the next subsection) in the formulation of diagnostic hypotheses, while a similar feature is not available for the specification of constrained latent classes.

BED, RS/KS, and DCM are rather different approaches. This complicates a straightforward comparison of these approaches. However, one thing is clear, in contrast to the other approaches BED does not need a calibration sample. This is a distinguishing feature of BED, it *can* be used if only the responses of the person to be diagnosed are available. However, note that there are situations where a calibration sample can be used to support the construction of informative diagnostic hypotheses. One was presented during the discussion of the dimensional perspective.

Another difference between BED and CLCM, RS/KS, and DCM is that BED uses the Bayes factor to evaluate hypotheses whereas the other approaches use data based weights and classification probabilities to assign persons to classes. The implication of this difference will further be elaborated in the last subsection of this section.

The Constructivist Perspective

In the constructivist perspective a diagnosis is based on the responses of a person to a convenient grouping of items (note that Borsboom, 2008, uses the word symptoms instead of items). The diagnostic and statistical manual of mental disorders (DSM-IV-TR, American Psychiatric Association, 2000) contains convenient groups of items that can be used to diagnose a person. One example are the seven items, for which abbreviations are displayed in Table 3, used for the diagnosis of an avoidant personality disorder. Scoring these items for a hypothetical person may render the response vector 1110011, where 1/0

denotes applies/does not apply to the person at hand.

Cognitive models can be based on existing guidelines (for example as in the DSM-IV-TR). These guidelines may or may not have been the result of calibration based on empirical research. According to the DSM-IV-TR an avoidant personality disorder is suspected if four or more items apply to the person at hand. This benchmark can, for example, be derived from a comparison of the number of items that apply between a group of healthy control persons and a group of persons suffering from avoidant personality disorder, that is, the benchmark is chosen such that the predictive validity of the test at hand is maximized.

This finding can be translated into hypotheses as follows:

$$H_{avoid} : \pi_j > 4/7 \text{ for } j = 1, \ldots, 7, \tag{16}$$

and

$$H_{non\ avoid} : \pi_j < 4/7 \text{ for } j = 1, \ldots, 7. \tag{17}$$

Note that, under H_{avoid} the expected number of applicable items is four or larger.

The Bayes factor comparing H_{avoid} and $H_{non\ avoid}$ is 17.00 for the response vector 1110011, that is, after observing the data the odds in favor of H_{avoid} have improved by 17.00, which in terms of the guidelines of Kass and Raftery (1995) constitutes positive evidence. As elaborated earlier in this paper, this conclusion only holds if a person's test taking behavior is as intended, that is, the hypotheses adequately represent the relevant cognitive models. It may, for example, be that the hypothesis displayed in Equation 16 is not a good representation of the relation between observed behavior (the item responses) and an avoidant personality disorder. An alternative representation will be given in the next subsection in which the causal systems perspective is introduced.

An exemplary other approach that fits within the constructivist perspective is mastery testing (MT; Dayton & Macready, 1988; Macready & Dayton, 1977, 1980). Mastery testing also involves a convenient grouping of items and a threshold that a person

has or has not crossed. In its simplest form mastery testing consists of three steps. In the first step the diagnostic test is provided to a sample of N persons from the population of interest. In the second step a restricted latent class model (McCutcheon, 1987; Vermunt & Magidson, 2005) is used to define masters (in the example at hand, those who have an avoidant personality disorder) and non-masters. Restrictions are applied such that the first class corresponds with a non-avoidant personality and the second class with an avoidant personality. It may, for example, be required that for each item the probability of the response "applies" is small in the first and large in the second class. In the third step the parameters of the latent class model are estimated, and for each person the probabilities of classification in the non-avoidant and avoidant class are computed. An elaboration of mastery testing can be found in Janssen, Tuerlinckx, Meulders, and De Boeck (2000) who use a hierarchical item response model and Glas and Vos (2010) and Vos and Glas (2010) who discuss adaptive mastery testing.

As presented in Table 4 within the constructivist perspective BED and MT differ in two ways. BED is more flexible than MT because: It requires only the responses of the persons to be diagnosed while MT requires a sample to calibrate a latent class model before a person can be diagnosed; as will be elaborated in the next two subsections, BED can use logical operators in the formulation of diagnostic hypotheses while MT can not use logical operators in the definition of latent classes; and, BED can use inequality constraints in the formulation of diagnostic hypotheses, while a similar feature is not available for the specification of latent classes in MT. Stated otherwise, within the constructivist perspective the scope of potential applications of BED is larger than the scope of applications of MT. There are also technical differences: BED uses the Bayes factor to evaluate hypotheses whereas MT uses data based weights and classification probabilities to assign persons to classes. The implication of these last two differences will be discussed in the last subsection of this section.

The Causal Systems Perspective

The fourth perspective is the causal systems perspective. In a causal system the probability of a positive response to the current item may be a function of the responses given to items preceding the current item in the causal system.

The causal systems perspective can also be applied to the items used to diagnose an avoidant personality disorder (see Borsboom, 2008, for another example involving the DSM-IV-TR). As is visualized in Figure 1, theoretical considerations may lead to a model stating that the roots of an avoidant personality disorder are person characteristics like: preoccupation with being criticized or rejected; feelings of social ineptness; and, reluctance to take social risks to avoid humiliation. The more of these items apply to a person, the higher the probability that a person is reluctant to participate in social involvement without assurance. If a person has this reluctance, the probability that occupational activities will be avoided, that inhibition in unfamiliar social situations occurs, and that fear of being shamed in close relationships occurs, will increase.

This hypothetical causal system can be formalized as follows:

$$H_{avoid}: \quad \begin{array}{c} \pi_4 > c, \pi_6 > c, \text{ and } \pi_7 > c \\ \text{and} \\ \text{if } S \geq 2 \text{ then } \pi_2 > c \\ \text{and} \\ \text{if } x_2 = 1 \text{ then } \pi_1 > c, \ \pi_3 > c, \text{ and } \pi_5 > c \end{array} , \tag{18}$$

and its complement

$$H_{non-avoid}: \text{not } H_{avoid}, \tag{19}$$

where S denotes the number of times *apply* is responded to items 4, 6, and 7. The "if" statements in Equation 18 are so-called logical operators. The first one states that, if the number of correct responses to items 4, 6, and 7 is larger or equal to two, then the probability of responding correctly to item two is larger than c. A further elaboration of

logical operators will be given in the next section. Like in the example presented for the diagnostic perspective, in this example a benchmark value c has to be specified. If a person has an avoidant personality disorder, it is clear that c should be a large probability, but it is unclear whether it should be .7, .8, or .9. A sensitivity analysis can be used to determine whether the choice among reasonable values for c really matters. For the response vector 1110011 the Bayes factor for the comparison of the hypotheses displayed in Equations 18 and 19 using values for c of .7, .8, and .9, renders values of 1.30, .79, and .25, respectively. This implies that for the response pattern at hand the conclusion *is* sensitive to the choice of c. According to the rules of Kass and Raftery (1995) the first two values are are not decisive, but .25 is positive evidence in favor of $H_{non-avoid}$. In such a case the optimal value for c can be determined using the predictive validity of informative diagnostic hypotheses specified using different values for c. However, if this is too costly or time consuming, a choice can only be based on expert judgement. Later in this paper it will be shown that for other response patterns the results are not sensitive to the choice of c.

Note that, the Bayes factor for the comparison of the hypotheses displayed in Equations 16 and 17 from the constructivist perspective was 17.00. This illustrates that the support in favor of hypotheses depends on the psychometric perspective chosen and the manner in which hypotheses are formalized. Consequently, it is very likely that the predictive validity of a test also depends on the manner in which the diagnostic hypotheses are formalized.

The causal systems perspective (Borsboom, 2008) is rather new, and has been used by Borsboom, Epskamp, Kievit, Cramer, and Schmittmann (2011), Cramer, Waldorp, van der Maas, and Borsboom (2010), and Schmittmann, Cramer, Waldorp, Epskamp, Kievit, and Borsboom (2011). There is not yet an established manner of data analysis within the causal systems perspective to which BED can be compared.

The General Form of Informative Diagnostic Hypotheses

Our comparison of BED with various other approaches was used to highlight the flexibility of BED compared to these other approaches. The goal is *not* to claim that one approach is better or superior to another. All are firmly established, and can be used without hesitation in the context for which they where developed. However, BED is a very general approach that can be used within each perspective. Furthermore, as exemplified in this section, there are examples of cognitive diagnostic assessment that can be handled by BED and not by the other approaches. This is not a fault of these other approaches. Each of them was designed for specific assessment problems and handles these problems appropriately. However, due to its flexibility BED provides a unifying framework for cognitive diagnostic assessment that encompasses all four psychometric perspectives.

The examples gave an impression of the informative diagnostic hypotheses that can be formulated and evaluated using BED. The basic form is:

$$H_m : \boldsymbol{R}_m\boldsymbol{\pi} > \boldsymbol{r}_m, \tag{20}$$

where: $\boldsymbol{R}_m$ is a matrix with J columns and K rows, where K denotes the number of restrictions imposed on $\boldsymbol{\pi}$; $\boldsymbol{\pi}$ denotes the unknown response probabilities for the person at hand; and $\boldsymbol{r}_m$ is a column with K constants. Note that the π's are not estimated, it is only determined which of several sets of constraints (the informative diagnostic hypotheses) superimposed on the π's is supported most by the response vector of the person to be diagnosed. Note that the hypothesis displayed in Equation 1 is obtained using

$$\boldsymbol{R}_m = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \tag{21}$$

and $\boldsymbol{r}_m = [0\ 0\ 0\ 0\ 0]$. The first line of $\boldsymbol{R}_m$ and the first number in $\boldsymbol{r}_m$ renders the

restriction $1 \times \pi_1 - 1 \times \pi_2 + 0 \times \pi_3 + 0 \times \pi_4 + 0 \times \pi_5 + 0 \times \pi_6 > 0$, that is, $\pi_1 > \pi_2$.

The general form of the informative diagnostic hypotheses that can be handled is is a combination of $p = 1, ..., P$ basic hypotheses that may or may not be active:

$$H_m : \text{If } f^p(\boldsymbol{x}) = 1 \text{ then } \boldsymbol{R}_m^p \boldsymbol{\pi} > \boldsymbol{r}_m^p, \text{ for } p = 1, ..., P. \tag{22}$$

The logical operator $f^p(\boldsymbol{x})$ is based on the item responses provided by the person at hand, if the logical operator equals 1/0 the corresponding inequality constraints are/are not evaluated. Note that the hypothesis displayed in Equation 18 is obtained using $P = 3$, $f^1(\boldsymbol{x}) = 1$ for each response vector, $f^2(\boldsymbol{x}) = 1$ if the number of times *apply* is responded to items 4, 6, and 7, is at least 2, and $f^3(\boldsymbol{x}) = 1$ if $x_2 = 1$. Furthermore,

$$\boldsymbol{R}_m^1 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{23}$$

and $\boldsymbol{r}_m^1 = [c\ c\ c]$. The first line of $\boldsymbol{R}_m^1$ and the first number in $\boldsymbol{r}_m^1$ render the restriction $0 \times \pi_1 + 0 \times \pi_2 + 0 \times \pi_3 + 1 \times \pi_4 + 0 \times \pi_5 + 0 \times \pi_6 + 0 \times \pi_7 > c$, that is, $\pi_4 > c$. Finally,

$$\boldsymbol{R}_m^2 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{24}$$

$\boldsymbol{r}_m^2 = [c]$,

$$\boldsymbol{R}_m^3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \tag{25}$$

and $\boldsymbol{r}_m^3 = [c\ c\ c]$.

Bayes Factor, p-values, and Classification Probabilities

On the one hand BED, MT, CLCA, RS/KS, DCM, and PFIT differ in the type of models and hypotheses that are used for cognitive diagnostic assessment. These differences have been highlighted in the previous subsections. On the other hand these approaches

differ in the statistical tool that is used to evaluate persons. If a latent class is seen as a hypothesis with respect to the process that generated a response vector, there are three main tools for the evaluation of hypotheses in the context of cognitive diagnostic assessment: Bayes factors, classification probabilities, and p-values.

There is by now a wealth of literature arguing in favor of Bayes factors and against the use of p-values for the evaluation of hypotheses. Among these are Morey and Rouder (2011), Rouder, Speckman, Sun, Morey, and Iverson (2009), Van de Schoot, Hoijtink, and Romeijn (2011), and Wagenmakers (2007). In the context of cognitive diagnostic assessment, the main point of critique with respect to the p-value is that it can not be used to quantify evidence in favor of the null-hypothesis. The implication is that p-value based person fit analysis as introduced under the dimensional perspective, cannot be used to provide evidence in favor of the null-hypothesis. This is a drawback because the research question is whether or not a person responds in agreement with an item response model. As elaborated earlier and also in the papers referred to above, the Bayes factor quantifies the support in the data for each hypothesis under consideration.

The Bayes factor and classification probabilities are closely related quantities. Since the prior distribution displayed in Equation 4 is constant, the Bayes factor for H_1 and H_2 can be written as:

$$BF_{12} = \frac{\int_{\boldsymbol{\pi} \in H_1} f(\boldsymbol{x}|\boldsymbol{\pi}) d\boldsymbol{\pi}}{\int_{\boldsymbol{\pi} \in H_2} f(\boldsymbol{x}|\boldsymbol{\pi}) d\boldsymbol{\pi}}. \tag{26}$$

If both H_1 and H_2 are point hypotheses, that is, $H_1 : \boldsymbol{\pi} = \boldsymbol{\pi}_1$ and $H_2 : \boldsymbol{\pi} = \boldsymbol{\pi}_2$, the Bayes factor reduces to:

$$BF_{12} = \frac{f(\boldsymbol{x}|\boldsymbol{\pi}_1)}{f(\boldsymbol{x}|\boldsymbol{\pi}_2)}. \tag{27}$$

The ratio of classification (RC) to class $q = 1$ and $q = 2$, respectively, is

$$RC_{12} = \frac{f(\boldsymbol{x}|\boldsymbol{\pi}_1)\omega_1}{f(\boldsymbol{x}|\boldsymbol{\pi}_2)\omega_2}, \tag{28}$$

where ω_1 and ω_2 denote the proportion of persons in the calibration sample assigned to class 1 and 2, respectively.

As can be seen, for the point hypotheses of interest Equations 27 and 28 are rather similar. The only difference is that the ratio of classification probabilities is also determined by the proportion of persons in the sample assigned to each of the classes. The Bayes factor can be transformed to posterior odds (the ratio of the posterior model probability of each hypothesis) if it is multiplied with prior odds (the ratio of the prior model probability of each hypothesis). For point hypotheses under consideration this renders:

$$\frac{PMP_1}{PMP_2} = \frac{f(\boldsymbol{x}|\boldsymbol{\pi}_1)P_1}{f(\boldsymbol{x}|\boldsymbol{\pi}_2)P_2}. \tag{29}$$

In fact, the Bayes factor is often interpreted as the posterior odds assuming that the prior odds are .5/.5, that is, equal prior probabilities for each of the hypotheses entertained. However, if a calibration sample is available, P_1 and P_2 could be estimated in which case the posterior odds and the ratio of classification probabilities turn out to be identical.

As was illustrated, for point null-hypotheses the Bayes factor and the ratio of classification probabilities are closely related. In general the Bayes factor as is used in this paper can be seen as a generalization of classification probabilities to a context in which the hypotheses of interest have been formulated using inequality constraints of the form presented in Equations 20 and 22.

Illustration of Diagnostic Testing

In this section the performance of the Bayes factor for the evaluation of the diagnostic hypotheses formulated earlier in this paper for each of the four perspectives on diagnostic testing will be illustrated. The effect of the number of items in a test on the size of the Bayes factor will be highlighted, and the (in)sensitivity to the choice of values for c, d, and e as appearing in Equations 13, 14, 15, 18, and 19 will be illustrated. Throughout this section it will be assumed that the response vectors to be evaluated resulted from persons who used one of the cognitive models that are covered by the informative diagnostic hypotheses under consideration.

The Dimensional Perspective

Table 6 consists of two panels. In the top panel, for five response vectors the support for the hypotheses

$$H_{cog} : \pi_1 > \pi_2 > \pi_3 > \pi_4 > \pi_5 > \pi_6 \tag{30}$$

and

$$H_{not\ cog} : \text{not } H_{cog}, \tag{31}$$

that were earlier in this paper introduced in the context of the dimensional perspective is displayed. Note that Equation 11 is used to compute the Bayes factor. Each response vector has the same number correct score, and the response patterns are ordered according to the number of Guttman errors G (Guttman, 1944; 1950). From IRT it is known that the larger the number of Guttman errors the worse the fit of a response pattern with an item response model.

As can be seen, as the number of Guttman errors increases the Bayes factor decreases, that is, the less a response vector is in agreement with an item response model, the smaller the support in the data for H_{cog}. As can be seen from the rapid change in the Bayes factor, it is about halved for each additional Guttman error, in a vector of six item responses each additional error contributes to a substantial decrease in the support for H_{cog}. Note that many person fit indices are based on or closely related to the number of Guttman errors in a response vector (Meijer & Sijtsma, 2001). As with the Bayes factor, it holds that the number of Guttman errors is inversely related to the fit of a response vector. It is therefore to be expected that the evaluation of a response vector using the Bayes factor will usually be in line with the evaluation by means of person fit indices. Another study like Karabatsos (2003) would be needed to properly compare the Bayes factor with these indices.

In the bottom panel it is assumed that a parallel test of six items has been added to the six items in the top panel, that is, each consecutive pair of items consists of two parallel items. Within the six response patterns displayed, the same response was given to

each pair of items. The following hypotheses that are generalizations of the hypotheses displayed in Equations (32) and (33) were evaluated:

$$H_{cog} : \{\pi_1, \pi_2\} > \{\pi_3, \pi_4\} > \{\pi_5, \pi_6\} > \{\pi_7, \pi_8\} > \{\pi_9, \pi_{10}\} > \{\pi_{11}, \pi_{12}\} \tag{32}$$

and

$$H_{not\ cog} : \text{not } H_{cog}. \tag{33}$$

As can be seen, the support for response vectors mostly in agreement with H_{cog} increases compared to the top panel (the first three), for response patterns in disagreement with H_{cog} the support decreases (the last one), and for response patterns that may or may not be in agreement with H_{cog} (the fifth) the Bayes factor remains almost unchanged compared to the corresponding pattern in the top panel and close to 1.0, that is, the value expressing no preference for either H_{cog} or $H_{not\ cog}$.

Finally note, that the Bayes factor as displayed in Equation 11 is computed using a sample of $\boldsymbol{\pi}$ from both its prior and posterior distribution. For all the analyses executed in this section the size of both samples is 10,000. This implies that there is Monte Carlo uncertainty in the estimates of f_m and c_m, and consequently in the estimate of the Bayes factor. Hoijtink (2012, Chapter 10) explains how a 95% interval reflecting the size of the Monte Carlo error can be computed. These intervals are displayed in Table 6. As can be seen, compared to the size of the Bayes factor these intervals are relatively small. However, if a larger accuracy is desired, this can be obtained using, for example, samples of 100,000 instead of 10,000.

The Diagnostic Perspective

The first panel of Table 7 contains three response vectors that are in decreasing degree in agreement with

$$H_{Rule1} : \pi_j > d \text{ for } j = 1, \ldots, 4, \text{ and } \pi_j < c \text{ for } j = 5, \ldots, 8, \tag{34}$$

which was presented in the discussion of the diagnostic perspective earlier in the paper. The second panel contains one response vector showing random response behavior as described by

$$H_{Rule2} : .33 - e < \pi_j < .33 + e \text{ for } j = 1, \ldots, 8, \tag{35}$$

and the third panel three response patters that are in decreasing degree in agreement with

$$H_{Rule3} : \pi_j > d \text{ for } j = 1, \ldots, 8. \tag{36}$$

In Table 8 a parallel test of eight items has been added to the eight items displayed in Table 7. Each consecutive pair of items consists of two parallel items to which the same response is given. The addition of items changes the hypotheses under consideration to:

$$H_{Rule1} : \pi_j > d \text{ for } j = 1, \ldots, 8, \text{ and } \pi_j < c \text{ for } j = 9, \ldots, 16, \tag{37}$$

$$H_{Rule2} : .33 - e < \pi_j < .33 + e \text{ for } j = 1, \ldots, 16, \tag{38}$$

and

$$H_{Rule3} : \pi_j > d \text{ for } j = 1, \ldots, 16. \tag{39}$$

As can be observed in each of the first three panels in Table 7, the Bayes factor comparing the correct rule with an unconstrained model is almost always larger than 1.0, that is, the constrained model is better than the unconstrained model, that is, the constrained model is a useful model in terms of Box's quote. This means that the response vector can be explained by the rule at hand. An exception can be observed for the seventh response vector, for which none of the Bayes factors is larger than 1.0. Apparently, this response vector is not well explained by any of the rules under investigation. Another observation is that the conclusions are not sensitive to the choice of c, d, and e.

The last two columns in Table 7 illustrate the gain that is obtained if competing hypotheses are compared, that is, if it is assumed that the rules under investigation adequately cover the response behavior of the persons to be evaluated. As can be seen, the mutual comparison of rules leads to more pronounced Bayes factors. This illustrates that

the specificity of the hypotheses to be compared influences the decisiveness of the resulting Bayes factors: The more specific the hypotheses, the more pronounced the Bayes factors. As is illustrated in Table 8, more pronounced Bayes factors will also be obtained if the number of items in a diagnostic test is increased.

The Constructivist Perspective

In Table 9 four response vectors are displayed that are decreasingly in agreement with

$$H_{avoid} : \pi_j > 4/7 \text{ for } j = 1, \ldots, 7, \tag{40}$$

and increasingly in agreement with

$$H_{non\ avoid} : \pi_j < 4/7 \text{ for } j = 1, \ldots, 7. \tag{41}$$

As can be seen, the performance of the Bayes factors is good. For response vectors with relatively many 1 responses the Bayes factor prefers H_{avoid} over an unconstrained hypothesis. The same holds for response vectors with relatively many 0 responses, but then with respect to $H_{non\ avoid}$. Furthermore, $BF_{avoid,non\ avoid}$ is positive to very strong for each response pattern, and decreasing from top to bottom.

The Causal Systems Perspective

In Tabel 10 three response patterns are displayed. The first is in agreement with

$$H_{avoid} : \begin{array}{c} \pi_4 > c, \pi_6 > c, \text{ and } \pi_7 > c \\ \text{and} \\ \text{if } S > 2 \text{ then } \pi_2 > c \\ \text{and} \\ \text{if } x_2 = 1 \text{ then } \pi_1 > c, \ \pi_3 > c, \text{ and } \pi_5 > c \end{array} \tag{42}$$

because two of items 4, 6, and 7, apply, item 2 applies, and each of items 1, 3, and 5 applies. The third is in agreement with

$$H_{non-avoid} : \text{not } H_{avoidant} \tag{43}$$

because only two items apply, and the second is less in agreement with H_{avoid} than the first but more than the third. Again the Bayes factors perform adequately. Going from the first to the third response vector the support in favor of H_{avoid} decreases. Furthermore, the results are not sensitive with respect to the choice of c. An exception may be the second response vector where values of c of .7 and .8 lead to indecisive Bayes factors, whereas $c = .9$ renders a Bayes factor that is positively in favor of $H_{non-avoid}$. As can finally be observed, using a sample of 10,000 from both the prior and posterior distribution of $\boldsymbol{\pi}$, the Monte Carlo error in the estimates of the Bayes factor is relatively small.

Discussion

This paper introduced Bayesian evaluation of informative diagnostic hypotheses with respect to the probability that a person will respond *yes* or *applies* to each of a set of items. As was illustrated in this paper, within each of four perspectives on diagnostic testing a flexible family of hypotheses can be specified. Furthermore, it was illustrated that evaluation of these hypotheses using the Bayes factor has good properties:

- If a response pattern is/is not in agreement with the various forms of hypotheses under consideration, the Bayes factor is usually positively larger than 3/smaller than .33. Furthermore the smaller the support in a response pattern for a hypothesis, the smaller the size of the corresponding Bayes factor. This could be observed in Tables 6 through 10
- Even if a test consists of only six to eight items, it is able to provide positive evidence for the hypotheses under consideration for a subset of the possible response patterns. As was shown using extensions with six items in Table 6 and eight items in Table 8, the larger the number of items in a test, the larger the evidence in favor or against the hypotheses under consideration.
- The result of a diagnostic test can be indecisive, that is, there is no forced decision in favor of one of the hypotheses under consideration if the amount of data needed for a decision is too small (see, for example, the second response vector in Table 10). If a

diagnostic test is indecisive, additional items can be provided to a person until a positive or strong result in favor of one of the hypotheses under consideration is obtained. This feature also opens the road to adaptive diagnostic testing.

- The specificity of the hypotheses under consideration is also of influence on the size of Bayes factor, that is, specific hypotheses are easier to evaluate than vague hypotheses. As can, for example, be seen in Table 8, for the response pattern 1100111111000000 the Bayes factor comparing $Rule1$ with an unconstrained hypothesis is 1.72, and thus not decisive. However, the comparison of this hypothesis with $Rule2$ gives a Bayes factors of 4.30 which is positive evidence in favor of $Rule1$.

Also, diagnostic testing using the Bayes factor is easy to implement using BED.exe (see the Appendix for further details). The result is a viable, practical, and applicable procedure that can be used for diagnostic testing.

This is probably not the last paper with respect to Bayesian evaluation of informative diagnostic hypotheses. Required and important contributions in this area would be applications of diagnostic testing including evaluations of the predictive validity of the hypotheses entertained for the population of interest. Furthermore, important elaborations of the approach presented in this paper would be adaptive diagnostic testing, that is, provide items to a person until the Bayes factor expresses strong evidence in favor of one of the hypotheses entertained; and repeated diagnostic testing, that is, evaluate a person at multiple time points, and base the diagnostic hypotheses on the data resulting from these multiple time points. Furthermore, this paper is limited to the evaluation of dichotomous item responses. However, applications in which polytomous or nominal item responses have to be evaluated are conceivable.

References

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders.* Washington, DC: American Psychiatric Association.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Boom, J., Hoijtink, H., & Kunnen, S. (2001). Rules in the balance. Classes, strategies or rules for the balance scale task. *Cognitive Development, 16, 717-735.*

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology, 64,* 1089-1108.

Borsboom, D, Epskamp, S., Kievit, R.A., Cramer, A.O.J, & Schmittmann, V.D. (2011). Transdiagnostic networks: commentary on Nolen-Hoeksema and Watkin (2011). *Perspectives on Psychological Science, 6,* 610-614.

Carmines, E.G., & Zeller, R.A. (1979). *Reliability and Validity.* London: SAGE.

Chiu, C., Douglas, J., & Li, X. (2009). Cluster analysis to cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633-655.

Cramer, A.O.J., Waldorp, O.J., van der Maas, H.L., & Borsboom, D. (2010). Comorbidity, a network perspective. *Behavioral and Brain Sciences, 33,* 137-193.

Dayton, M., & MacReady, G.B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association, 83,* 173-178.

De Boeck, P., & Wilson, M. (2004). *Explanatory Item Response Models: a Generalized Linear and Nonlinear Approach.* New-York, NJ: Springer.

De Carlo, L.T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36,* 447-468.

De La Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 2,* 179-199.

DiBello, L.V., Roussos, L.A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In: C.R. Rao & S. Sinharay (Eds.), *Handbook of Statistics, Volume 26: Psychometrics*, pp. 979-1030, Amsterdam : Elsevier B.V.

Doignon, J.P., & Falmagne, J.CI. (1999). *Knowledge Spaces.* New York: Springer.

Embretson, S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists.* New York, NJ: Lawrence Erlbaum.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Fischer, G.H., & Molenaar, I.W. (1995). *Rasch Models, Foundations, Recent Developments and Applications.* New York, NJ: Springer.

Fox, J-P (2010). *Bayesian Item Response Theory Modeling: Theory and Applications.* New York, NJ: Springer.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian Data Analysis.* Boca Raton, FL: Chapman and Hall/CRC.

Glas, C.A.W., & Vos, H.J. (2010). Adaptive mastery testing using a multidimensional IRT model. In: W.J. van der Linden & C.A.W. Glas (Eds.), *Elements of Adaptive Testing.* New York: Springer.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139-150.

Guttman, L. (1950). The basis for scalogram analysis. In: S.A. Stouffer, L.Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, & J.A. Claussen (Eds.), *Measurement and Prediction*, pp. 60-90. Princeton, NJ: Princeton University Press.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory. Principles and Applications.* Boston, MA: Kluwer.

Hoijtink, H. (2001). Confirmatory latent class analysis: model selection using Bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research, 36,* 563-588.

Hoijtink, H. (2012). Informative Hypotheses: Theory and practice for behavioural and social scientists. Boca Raton, FL: Chapman and Hall/CRC.

Hoijtink, H., & Boom, J. (2008). Inequality constrained latent class analysis. In: H. Hoijtink, I. Klugkist, & P.A. Boelen, *Bayesian Evaluation of Informative Hypotheses*, pp. 227-246. New York, NJ: Springer.

Jacob, B.A., & Levitt S.D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118,* 843-877.

Jaeger, R.M. (1988). Use and effect of caution indices in detecting aberrant patterns of standard setting judgements. *Applied Measurement in Education, 1,* 17-31.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25*, 285-306.

Jeffreys, H. (1961). *Theory of Probability. Third Edition.* Oxford: Oxford University Press.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty six person fit statistics. *Applied Measurement in Education 16,* 277-298.

Kass, R.E., & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

Klauer, K.C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56,* 213-228.

Klauer, K.C. (1995). The assessment of person fit. In: G.H. Fischer & I.W. Molenaar (Eds.). *Rasch Models*, pp. 97-110. New York: Springer.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality Constrained Analysis of Variance: A Bayesian approach. *Psychological Methods, 10,* 477-493.

Kripke, S.A. (1982). *Wittgenstein on Rules and Language.* Cambridge, MA: Harvard University Press.

Laudy, O., Boom, J., & Hoijtink, H. (2004). Bayesian computational methods for inequality constrained latent class analysis. In: A. van der Ark, M. Croon, & K. Sijtsma (Edss). *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences.* Mahwah, NJ: Erlbaum.

Lavine, M., & Schervish, M.J. (1999). Bayes factors: What they are and what they are not. *The American Statistician, 53,* 119-122

Leighton, J.P., & Gierl, M.J. (2007). Cognitive Diagnostic Assessment for Education. Theory and Applications. Cambridge: Cambridge University Press.

Liu, J., Xu, G., & Ying, Z. (2012). Data driven learning of Q-matrix. *Applied Psychological Measurement, 36,* 609-618.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Erlbaum.

Lynch, S.M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists.* New York, NJ: Springer.

Macready, G.B., & Dayton, M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*, 99-120.

Macready, G.B., & Dayton, M. (1980). The nature and use of state mastery models. *Applied Psychological Measurement, 4*, 493-516.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212.

McCutcheon, A.L. (1987). *Latent Class Analysis.* Thousand Oaks, California: SAGE.

Meehl, P.E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194-216.

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

Meijer, R.R., Egberink, I.J.L., Emons, W.H.M., & Sijtsma, K. (2008). Deterction and validation of unscalable item score patterns using item response theory: An illustration with Harter s self-perception profile for children. *Journal of Personality Assessment, 90,* 227-238.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor Approaches for Testing Interval Null Hypotheses. *Psychological Methods, 16,* 406-419.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference, 140,* 887-906.

Rijkes, C.P.M., & Kelderman, H. (2006). Latent-response loglinear-Rasch models for strategy shifts in problem solving processes. In: M. von Davier & C.H. Cartstensen (Eds.), Multivariate and Mixture Distribution Rasch Models. New York: Springer.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review, 16,* 225-237.

Rupp, A.A., Templin, J., & Henson, R.A. (2010). Diagnostic Measurement. Theory, Methods, and Applications. New York: The Guilford Press.

Schmittmann, V.D., Cramer, A.O.J., Waldorp, O.J., Epskamp, S., Kievit, R.A., & Borsboom, D. (2011). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology.* doi:10.1016/j.newideapsych.2011.02.007

Schonemann, P.H., & Thompson, W.W. (1996). Hit rate bias in mental testing. *Cahiers de Psychologie Cognitive, 15,* 2-28.

Schonemann, P.H. (1997). Some new results on hit rates and base rates in mental testing. *Chinese Journal of Psychology, 39*, 173-192.

Schonemann, P.H. (2005). Psychometrics of Intelligence. In: K. Kempf-Leonard, *Encyclopedia of Social Measurement, Volume 3.* Amsterdam: Elsevier.

Schrepp, M. (2005). About the connection between knowledge structures and latent class models. *Methodology, 1,* 92-102.

Siegler, R.S. (1981). *Developmental sequences within and between concepts.* Monographs of the Society for Research in Child Development, 46, 2 (Serial No. 189).

Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: SAGE.

Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Tatsuoka, K.K. (2009). *Cognitive Assessment: An Introduction to the Rule Space Method.* New York: Routledge.

Tatsuoka, K.K., & Tatsuoka, M.M. (1997). Computerized cognitive diagnostic adaptive testing: Effects on remedial instruction as empirical validation. *Journal of Educational Measurement, 34*, 3-20.

Taylor, H.C., & Russell, J.T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology, 23,* 565-578.

Van de Schoot, R., Hoijtink, H., & Romeijn, J.W. (2011). Moving beyond traditional null hypothesis testing: Evaluating expectations directly. *Frontiers in Psychology, 2,* Article 24. doi:10.3389/fpsyg.2011.00024

Vermunt, J.K., & Magidson J. (2005). *Technical Guide for Latent GOLD 4.0: Basic and Advanced*. Belmont Massachusetts: Statistical Innovations Inc.

Von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classifications models. *Measurement: Interdiciplinary Research and Perspective, 7*, 67-74.

Von Davier, M. (2011). Equivalence of the DINA model and a constrained general diagnostic model. Educational Testing Service, Research Report RR-11-37. Retrieved from http://www.ets.org/Media/Research/pdf/RR-11-37.pdf

Vos, H.J., & Glas, C.A.W. (2010). Testlet based adaptive mastery testing. In: W.J. van der Linden & C.A.W. Glas (Eds.), *Elements of Adaptive Testing.* New York: Springer.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review, 14*, 779-804.

Wang, C., Chang, M., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods, 44,* 95-109.

Table 1

The Context for Cognitive Diagnostic Assessment

	Cognitive Domain of Interest		
Psychometric Perspective	Arithmetic Ability	Developmental Stages	Avoidant Personality
Dimensional	X		
Diagnostic		X	
Constructivist			X
Causal System			X

Table 2

True and Predicted Diagnosis

	Predicted	
True	m	m'
m	T_m	$F_{m'}$
m'	F_m	$T_{m'}$

Table 3

Diagnosis of Avoidant Personality Disorder

Item	Abbreviated Phrasing
1	Avoids occupational activities that require interpersonal contact
2	Reluctant to participate in social involvement without assurance
3	Fears of being shamed or ridiculed in close relationships
4	Preoccupied with being criticized or rejected
5	Inhibited in unfamiliar social situations due to feelings of inadequacy
6	Regards him- or herself as socially inept
7	Reluctant to take social risks in order to avoid possible humiliation

Table 4

Comparison of Various Approaches

Property	BED	PFIT	MT	CLCA	RS/KS	DCM
Calibration Sample	N	Y	Y	Y	Y	Y
Logical Operators	Y	N	N	N	N	N
Inequality Constraints	Y	N	N	Y	N	N
Weights	N	NA	Y	Y	Y	Y
Diagnostic Tool	BF	p-value	CP	CP	CP	CP

Note. N=No, Y=Yes, BF=Bayes Factor, CP=Classification Probabilities, NA=Not Applicable.

Table 5

Balance Scale Items

Item	Left Side		Right Side		Item Type	
	Dist.	Weight	Dist.	Weight	Conflict Weight	Conflict Balance
1	2	3	4	1	1	0
2	2	4	4	1	1	0
3	3	1	1	4	1	0
4	3	2	4	1	1	0
5	1	3	3	1	0	1
6	1	4	2	2	0	1
7	3	2	2	3	0	1
8	4	1	2	2	0	1

Table 6

Bayes Factors for Hypotheses from the Dimensional Perspective

G	$\boldsymbol{x}$	$BF_{cog,notcog}$	Lower Bound - Upper Bound
0	111000	7.76	7.00-8.44
1	110100	3.91	3.52-4.35
2	101100	1.82	1.64-2.03
3	101010	.85	.74-.96
4	100110	.63	.55-.72
	111111000000	83.56	72.57-96.10
	111100110000	18.73	16.04-21.86
	110011110000	4.48	3.83-5.26
	110011001100	.99	.82-1.19
	110000111100	.42	.35-.51

Table 7

Bayes Factors for Hypotheses from the Diagnostic Perspective

$\boldsymbol{x}$	$BF_{Rule1,u}$	$BF_{Rule2,u}$	$BF_{Rule3,u}$	$BF_{Rule1,Rule2}$	$BF_{Rule1,Rule3}$
11110000	96.18	.54	.015	178.11	6412.00
	(142.76/79.33)	(.64)	(.001/.03)		
10110000	11.58	1.17	.001	9.89	11580.00
	(8.18/12.17)	(1.30)	(.000/.005)		
10111000	1.24	.55	.018	2.25	68.88
	(.37/1,68)	(.63)	(.001,.04)		
$\boldsymbol{x}$	$BF_{Rule1,u}$	$BF_{Rule2,u}$	$BF_{Rule3,u}$	$BF_{Rule2,Rule1}$	$BF_{Rule2,Rule3}$
10100100	.14	1.19	.001	8.50	1190.00
	(.03,.25)	(1.36)	(.000,.005)		
$\boldsymbol{x}$	$BF_{Rule1,u}$	$BF_{Rule2,u}$	$BF_{Rule3,u}$	$BF_{Rule3,Rule1}$	$BF_{Rule3,Rule2}$
11111111	.014	.029	100.82	7201.42	3476.55
	(.001/.03)	(.036)	(163.81/85.20)		
10111011	.015	.13	1.31	87.33	10.07
	(.001/.03)	(.14)	(.46,1.85)		
10011011	.001	.28	.14	140.00	.50
	(.000,.005)	(.30)	(.02/.25)		

Note. First line $d = .8, c = .2, e = .05$; second line: Rule 1

$d = .9, c = .1/d = .75, c = .25$; Rule 2, $e = .10$; Rule 3 $d = .9/d = .75$.

Table 8

Bayes Factors for Hypotheses from the Diagnostic Perspective

$\boldsymbol{x}$	$BF_{Rule1,u}$	$BF_{Rule2,u}$	$BF_{Rule3,u}$	$BF_{Rule1,Rule2}$	$BF_{Rule1,Rule3}$
1111111100000000	11702.10	.40	.0002	>20000	>20000
1100111100000000	135.26	1.66	.000002	81.48	>20000
1100111111000000	1.72	.40	.0002	4.30	8600.00
$\boldsymbol{x}$	$BF_{Rule1,u}$	$BF_{Rule2,u}$	$BF_{Rule3,u}$	$BF_{Rule2,Rule1}$	$BF_{Rule2,Rule3}$
1100110000110000	.02	1.64	.000003	82.00	>20000
$\boldsymbol{x}$	$BF_{Rule1,u}$	$BF_{Rule2,u}$	$BF_{Rule3,u}$	$BF_{Rule3,Rule1}$	$BF_{Rule3,Rule2}$
1111111111111111	.0002	.001	11789.56	>20000	>20000
1100111111001111	.0002	.02	1.64	8200.00	82.00
1100001111001111	.000003	.10	.02	6666.66	.2

Note. $d = .8, c = .2, e = .05$.

Table 9

Bayes Factors for Hypotheses from the Constructivist Perspective

$\boldsymbol{x}$	$BF_{avoid,u}$	$BF_{non\ avoid,u}$	$BF_{avoid,non\ avoid}$
1111111	24.78	.02	1239.00
1111100	2.04	.12	17.00
1110000	.13	.75	.17
1000000	.01	4.65	.002

Table 10

Bayes Factors for Hypotheses from the Causal Systems Perspective

$\boldsymbol{x}$	c	$BF_{avoid,c}$	Lower Bound - Upper Bound
1111110	.7	6.84	6.11-7.65
	.8	6.40	5.47-7.48
	.9	5.70	4.39-7.43
1111010	.7	1.23	1.08-1.39
	.8	.73	.61-.88
	.9	.30	.21-.41
0101000	.7	.00	.00-.00
	.8	.00	.00-.00
	.9	.00	.00-.00

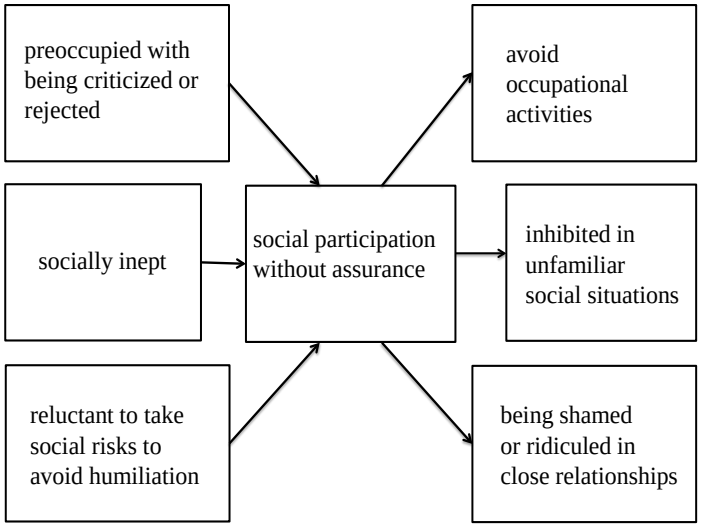

Figure 1. A Causal System for the Avoidant Personality Disorder Data

Appendix

Implementation in BED.exe

All the analyses presented in this paper can be executed using BED.exe. It estimates Bayes factors and provides a 95% Monte Carlo interval based on the decomposition presented in Chapter 10 of Hoijtink (2012). A zip-file containing BED.exe can be downloaded from http://tinyurl.com/hoijtinkbook under the rubric "New Developments and Software". It further contains two support files, that are not of further importance to users, and an input file named "input.txt". The input file is a text file that can be modified as required. An illustrative example is presented below.

```
Nitem Nrestr Niter
   7       7    10000
R-r
0 0 0 1 0 0 0 .7
0 0 0 0 1 0 0 .7
0 0 0 0 0 1 0 .7
0 1 0 0 0 0 0 .7
1 0 0 0 0 0 0 .7
0 0 1 0 0 0 0 .7
0 0 0 0 1 0 0 .7
initpi
.9 .9 .9 .9 .9 .9 .9
response vector
1 1 1 1 0 1 0
```

The first line contains labels. On the second line below "Nitem" the number of items has to be recorded, below "Nrestr" the number of active restrictions used to specify a

hypothesis, and below "Niter" the number 10,000 (this number will further be discussed below). Note that all numbers in this file have to be separated by spaces.

If a logical operator is used to specify an hypothesis, a restriction is only active if the logical operator is true for the response vector to be evaluated (recorded in the last line of the input file). Only active restrictions are included below the label "R-r". What is recorded are the matrices $\boldsymbol{R}_m^p$, and the vectors $\boldsymbol{r}_m^p$ from Equation 22 for the active restrictions. The first "Nitem" columns belong to $\boldsymbol{R}$ and the last column belongs to $\boldsymbol{r}$. The general structure of each line is $R_1, \ldots, R_J, r$, which renders the restriction $R_1\pi_1 + ... + R_J\pi_J > r$. To elaborate the meaning of the lines following the label "R-r" a number of examples will be given:

- 1 -1 0 0 0 denotes that $\pi_1 - \pi_2 > 0$, that is, $\pi_1 > \pi_2$
- -1 1 0 0 0 denotes that $-\pi_1 + \pi_2 > 0$, that is, $\pi_2 > \pi_1$
- 0 0 1 0 .8 denotes that $\pi_3 > .8$
- 0 0 -1 0 -.2 denotes that $-\pi_3 > -.2$, that is, $\pi_3 < .2$
- 2 -1 -1 0 .5 denotes that $2\pi_1 - \pi_2 - \pi_3 > .5$

Below the label "initpi" values for the π's are given that are in agreement with the hypothesis specified below the label "R-r". Usually there are many options, any one of these options will do. Below the label "response vector" the item responses of the person to be diagnosed are presented.

The example input file corresponds to H_{avoid} from the causal systems perspective displayed in Equation 18:

$$H_{avoid}: \quad \begin{array}{c} \pi_4 > .7, \pi_6 > .7, \text{ and } \pi_7 > .7 \\ \text{and} \\ \text{if } S \geq 2 \text{ then } \pi_2 > .7 \\ \text{and} \\ \text{if } x_2 = 1 \text{ then } \pi_1 > .7, \ \pi_3 > .7, \text{ and } \pi_5 > .7. \end{array} \quad . \tag{44}$$

For this hypothesis "Nitem=7". Looking at the response vector 1111010 presented on the

last line of the input file, it can be seen that all the restrictions are active, that is, "Nrestr=7". These restrictions are presented below the line labeled "R-r". The values for the π's recorded below the label "initpi" are in agreement with these restrictions.

Running the example input file renders the following output file:

fit =
0.0002695791

complexity =
0.0002189452

Bayes factor for hypothesis versus unconstrained alternative =
1.2312630000
interval for Bayes factor due to estimation error =
1.0853211000 1.3976720000

Bayes factor for hypothesis versus complement =
1.2313253000
interval for Bayes factor due to estimation error =
1.0853420000 1.3977874000

The estimated complexity, fit, and Bayes factor for the hypothesis at hand compared to an unconstrained hypothesis and its complement are displayed. For both Bayes factors a 95% Monte Carlo interval reflecting the uncertainty in the estimates is provided. If the interval is considered to be too large compared to the size of the Bayes factor, the size of "Niter" has to be increased. For all the computations presented in this paper "Niter=10000". This rendered reasonably small Monte Carlo intervals. The accuracy of the

estimates can be increased using "Niter=100000" or even larger values.

To compare an hypothesis H_m to one or more competing hypotheses (see, for example, Equations 13, 14, and 15), BED.exe must be run for each hypothesis under investigation. This will render BF_{mu} for each hypothesis under consideration. Subsequently Equation 10 can be used to compare each pair of hypotheses.