A causal theory of error scores

van Bork, R.; Rhemtulla, M.; Sijtsma, K.; Borsboom, D.

DOI
10.1037/met0000521
Publication date
2024
Document Version
Author accepted manuscript
Published in
Psychological Methods

A Causal Theory of Error Scores

Article *in* Psychological Methods · July 2022

DOI: 10.1037/met0000521

4 authors, including:

Riet van Bork
University of Amsterdam
38 PUBLICATIONS **9,159** CITATIONS
SEE PROFILE

Mijke Rhemtulla
University of California, Davis
87 PUBLICATIONS **8,949** CITATIONS
SEE PROFILE

Denny Borsboom
University of Amsterdam
374 PUBLICATIONS **55,793** CITATIONS
SEE PROFILE

A Causal Theory of Error Scores

Riet van Bork[1,2], Mijke Rhemtulla[3], Klaas Sijtsma[4], and Denny Borsboom[1]

[1]Department of Psychology, University of Amsterdam
[2]The Center for Philosophy of Science, University of Pittsburgh
[3]Department of Psychology, University of California, Davis
[4]Department of Methodology and Statistics, Tilburg University

Abstract

In Modern Test Theory, response variables are a function of a common latent variable that represents the measured attribute, and error variables that are unique to the response variables. While considerable thought goes into the interpretation of latent variables in these models (e.g., validity research), the interpretation of error variables is typically left implicit (e.g., describing error variables as residuals). Yet, many psychometric assumptions are essentially assumptions about error and thus being able to reason about psychometric models requires the ability to reason about errors. We propose a causal theory of error as a framework that enables researchers to reason about errors in terms of the data-generating mechanism. In this framework, the error variable reflects myriad causes that are specific to an item and, together with the latent variable, determine the scores on that item. We distinguish two types of item-specific causes: characteristic variables that differ between people (e.g., familiarity with words used in the item), and circumstance variables that vary over occasions in which the item is administered (e.g., a distracting noise). We show that different assumptions about these unique causes (1) imply different psychometric models, (2) have different implications for the chance experiment that makes these models probabilistic models, and (3) have different consequences for item bias, local homogeneity, and reliability coefficient α and the test-retest correlation. The ability to reason about the causes that produce error variance puts researchers in a better position to motivate modeling choices.

Keywords: causal theory of error, item response theory, error definition of IRT models, latent variable models

This version is the final pre-formatted version that has been accepted for publication in *Psychological Methods*, see: https://doi.org/10.1037/met0000521.

Correspondence concerning this article should be addressed to Riet van Bork, Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129B, 1018 WT Amsterdam, the Netherlands. Email: rietvanbork@hotmail.com.

Acknowledgements: We thank the Pittsburgh spring 2021 fellows reading group (Edouard Machery, Ryan Nefdt, Mike Schneider, Hannah Rubin, Christopher Weaver and Nicholas Huggett) for input on an earlier draft of this paper. We also thank Paul de Boeck, Willem Heiser and Lourens Waldorp for conversations that contributed to the revisions, and several anonymous reviewers for their thoughtful comments that helped improve this article.
Funding: This work was supported by Consolidator Grant 647209 of the European Research Council (ERC) and by Grant VI.C.181.029 of the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).
The ideas in this manuscript have been presented at the 84th Annual Meeting of the Psychometric Society (IMPS) in 2019, and a preprint of an earlier version of this manuscript was shared on PsyArxiv.

1 Introduction

A person who takes an intelligence test the day after a night of having slept badly might claim that their score does not reflect their intelligence because they were tired. Likewise, a person experiencing a concentration lapse during a test, may find that a cup of coffee can improve their test score. And, for a person who is unfamiliar with some of the words used to introduce the problems in a math test, the resulting test score is not a valid assessment of their math ability. Such claims, which are commonplace in the discourse concerning psychological test scores, stem from intuitive ideas about the causal background of test scores, and can be phrased in terms of counterfactuals like "If I had done this same test after a good night's sleep I would have scored higher", "If I had not had coffee before the test I would have scored lower" and "If I had understood the words in the introductory story, I would have solved the math problem".

Such arguments rest on the idea that item scores are partly determined by factors unrelated to the measured attribute, such as fatigue, concentration level or familiarity with the vocabulary (Cronbach, 1971). Some of these factors cause variation of an individual's score across testing occasions given a fixed attribute level (e.g., temporary changes in energy level as a result of the amount of sleep or a cup of coffee), and some result in variation in scores between people or groups of people who have the same attribute level (e.g., differences in familiarity with words affecting scores on a math test). In latent variable models, both types of factors result in residual variance and are typically referred to as random measurement error and systematic error, respectively. These errors cause different psychometric phenomena: due to random measurement error, item scores are not perfectly *reliable* (Lord & Novick, 1968; Mellenbergh, 1996) and due to systematic error, item scores are *biased* with respect to the measured attribute (Mellenbergh, 1989; Meredith, 1993; Millsap, 1997, 2007).

The examples illustrate an intuition about what constitutes error. Despite an intuitive understanding of which factors qualify as error, a substantive interpretation of the error scores in a statistical model is typically left implicit. That is, while most variables in the model are interpreted as reflecting meaningful constructs, errors are commonly described as *residuals*, the remaining variability that is not accounted for by the rest of the model (Kline, 2016, p. 13; Newsom, 2015, p. 2; Raykov & Marcoulides, 2000, p. 13). The lack of interpretation of error is at odds with the importance of the role that error has in psychometric models. Error is what makes psychometric models probabilistic, and all quantitative information (going beyond merely ordinal) is derived from the distribution of error (Michell, 2004). Error is where the action is!

We restrict our analysis to a fixed set of items [1] and to the unidimensional latent variable model that is characterized by the assumptions of local independence, unidimensionality, and monotonicity. We turn this model upside down to build a causal theory of error in which the error variable is a composite of causal effects on the item score that are unique to that item. This theory builds on reflective measurement theory in which item scores are *effects* of the latent variable (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000). The unique causes of the item score are not different from the latent variable except that they are item-specific. Some causes vary over measurement occasions, resulting in random measurement error, and some causes are variables on which people differ systematically, resulting in systematic error. As a result, the distinction between latent variables and error no longer distinguishes substantive variables from noise. We view the response variable as the effect of myriad causal factors that can be distinguished on two relevant dimensions: (1) Causes can be *shared* among multiple items, or they can be *unique* to an item. Common factors are of the first kind and error variables of the second. (2) Causes can vary over *persons*, and over *measurement occasions*. One arrives at other latent variable models by adding or relaxing assumptions about these causes of the response variable.

We show that many assumptions of psychometric models are essentially assumptions about error. Hence, reasoning about psychometric models is reasoning about error. The causal theory of error that we propose in this article serves as a theoretical framework that helps reasoning about error and provides a causal interpretation of assumptions in psychometric models that we explicate in this paper.

In the next section, we first introduce the random variables and psychometric concepts that are needed to define our causal theory of error. We then present a causal theory of error in which the error variables in latent variable models reflect a multitude of causal effects that are specific to

[1] By restricting our analysis to a fixed set of items, we consider the random sampling of people and of occasions but not of items.

the response variable. In the sections thereafter, we show how assumptions about error relate to (1) assumptions of different latent variable models, (2) the chance experiment that underlies the randomness in item responses, and (3) item bias, local homogeneity, and reliability methods such as coefficient α and the test-retest correlation. By giving an interpretation to these assumptions in terms of the data-generating mechanism, the proposed causal theory of error helps reasoning about these assumptions.

2 Test Theory

Let $\mathbf{X} = [X_1, \dots X_J]$ denote a vector of J response variables, in which X_j denotes a random variable with scores on item j, of which the sample space is defined by the random selection of people from a population (Lord & Novick, 1968, pp. 32–34)[2]. For example, $\mathbf{X}$ could represent the scores on several math problems, of which X_j represents the scores on one such math problem, for example, '$12 + 4 = \dots$'.

In Classical Test Theory (CTT), a specific person i has a true score for item j which is operationally defined as the expected value of person i's scores to item j over repeated measures in a hypothetical experiment in which the person is "brainwashed" to remove their memory of responding and presented with the same item repeatedly (Lazarsfeld, 1959; quoted in Lord & Novick, 1968, p. 30),

$$\tau_{ij} = E(X_{ij}). \tag{1}$$

The probability distribution of X_{ij} is what Lord and Novick (1968, p. 47) call a *propensity distribution*, and X_{ij} derives its randomness from the random sampling of measurement occasions from the set of occasions in which person i could have responded to item j. Measurement error is operationally defined as the difference between the observed item score and the true score,

$$E_{ij}^{CTT} = X_{ij} - \tau_{ij}, \tag{2}$$

where E_{ij}^{CTT}, like X_{ij}, is a random variable for a specific person i defined over hypothetical measurement occasions. We use the superscript CTT to contrast this error variable to the error variable, E_j, in Modern Test Theory that we introduce shortly and is the focus of this article. By randomly selecting people from a population, response variable X_j, true-score variable T_j, and measurement-error variable E_j^{CTT} are obtained, of which the latter is the difference between the first two (Lord & Novick, 1968, pp. 32–34),

$$E_j^{CTT} = X_j - T_j. \tag{3}$$

In the remainder of this article, we use *occasion* to refer to a constellation of uncontrolled factors that form hypothetical conditions under which the item can be administered (Ellis & Van den Wollenberg, 1993). Let $\mathcal{P}$ be a set of subjects, and $\mathcal{K}$ a set of occasions. The subscript i is important, because it distinguishes X_{ij}, which is a variable for a specific person i of which the sample space is defined only by the random selection of occasions from $\mathcal{K}$, from X_j, which is a variable for which the sample space is defined by selecting pairs of a subject and an occasion from the domain $\mathcal{P} \times \mathcal{K}$. The response variable X_j thus has two sources of probability; the sampling of people from a population, and the sampling of a measurement occasion for each sampled person. These two dimensions are key to the two types of causes in the causal theory of error that we define in the next section.

The distinction between variability over measurement occasions and over persons is also at the heart of the concept of reliability (Cronbach, 1947; Lord & Novick, 1968; Miller, 1995; Sijtsma & Van der Ark, 2015). The reliability of X_j is defined as the proportion of observed-score variance of X_j that is true-score variance,

$$\rho_{X_j X_j'} = \frac{\sigma_{T_j}^2}{\sigma_{X_j}^2} = 1 - \frac{\sigma_{E_j^{CTT}}^2}{\sigma_{X_j}^2}. \tag{4}$$

When X_j is perfectly reliable (i.e., reliability equal to 1), all variability in X_j is variability in the true scores between persons and given a specific person i, x_{ij} is a constant. That is, $X_j = T_j$,

[2] Lord and Novick(1968, pp. 32-34) use the subscript $*$ to denote the random selection of people from a population: X_{j*}. This differs from the subscript i, which denotes the selection of a specific person i. We leave out the subscript $*$.

and because T_j only varies over persons, any perfectly reliable variable has the sampling of people from $\mathcal{P}$ as its only source of probability. When X_j is perfectly unreliable (i.e., reliability 0), people have the same true score and all variance in X_j reflects variability over measurement occasions. Equations (1) and (2) imply that all people have an expected error score of 0. All variability in the error variable E_j^{CTT} in Equation (3) is variability over measurement occasions and E_j^{CTT} has reliability 0 by definition. In the following, we restrict our use of the term *measurement error* only to this type of perfectly unreliable (random) error. The error variable in CTT thus only comprises measurement error.

Modern Test Theory (MTT) is a test-theoretical framework in which latent variable models (Junker & Ellis, 1997; Mellenbergh, 1994a; Moustaki & Knott, 2000) are used to analyze and construct tests that measure mental attributes. A vector of response variables $\mathbf{X}$ is regressed on one or more latent variables Θ that refer to psychological attributes (e.g., math ability) on which people show systematic variation. We focus on the situation where Θ is a single unidimensional latent variable. Most of the important latent variable models are specializations of

$$X_j = f(\Theta) + E_j, \tag{5}$$

by (a) substituting appropriate distribution functions for $\mathbf{X}$ and Θ, and (b) choosing the appropriate Item Response Function (IRF; Borsboom & Molenaar, 2015; Irwing, Booth, & Hughes, 2018; Mellenbergh, 1994a; Sijtsma & Van der Ark, 2020) to relate $\mathbf{X}$ to Θ. For example, in linear Confirmatory Factor Analysis (CFA) models (e.g., Bartholomew & Knott, 1999; Bollen, 1989; Jöreskog, 1971; Sörbom, 1974; Spearman, 1904), $\mathbf{X}$ consists of continuous variables that are normally distributed, and the IRF that relates these variables to Θ is linear,

$$X_j = \nu_j + \lambda_j \Theta + E_j. \tag{6}$$

Here λ_j denotes the loading of item j on the latent variable, and ν_j denotes the intercept (Jöreskog, 1971). In the Item Response Theory (IRT) models (e.g., Birnbaum, 1968; Goldstein & Wood, 1989; Guttman, 1950; Lord, 1952, 1980; Mellenbergh, 1994a; Mokken, 1971; Moustaki & Knott, 2000; Rasch, 1960; Thissen & Steinberg, 1986; Van der Linden & Hambleton, 1997) we discuss in this article, $\mathbf{X}$ consists of binary variables that are Bernoulli distributed and the IRF that relates $\mathbf{X}$ to Θ is logistic or cumulative normal. These models are probabilistic models of the general form in Equation (5) in which E_j accounts for residual variation in X_j given $\Theta = \theta$.

In CTT, E_j^{CTT} is defined as perfectly unreliable. In MTT, E_j can be perfectly unreliable but is not necessarily so. In contrast to the true score variable in CTT, Θ represents not just all reliable predictors of the item scores but reflects a dimension that renders the set of response variables conditionally independent (Ellis & Junker, 1997). We focus here on a subset of commonly used latent variable models that are characterized by three assumptions: (1) Θ is unidimensional, (2) $\mathbf{X}$ is conditionally independent given Θ (i.e., local independence) and (3) $P(X_j > x \mid \Theta)$ is monotonically nondecreasing in Θ. These assumptions allow for E_j in Equation (5) to have a reliability larger than zero as long as all variance in E_j is unique to item j. The reliable part of the error variable is referred to as *systematic error* and the unreliable part as *measurement error* or *random error*.

This section has introduced the sampling from the domain of people and from the domain of occasions as two sources of probability for response variables. The causal theory of error that is presented in the next section builds on these two domains and on the distinction between perfectly reliable variables and perfectly unreliable variables. For perfectly reliable variables, the scores are constant across occasions in $\mathcal{K}$ and only vary between people in $\mathcal{P}$. For perfectly unreliable variables, the scores vary over occasions in $\mathcal{K}$, and the expected value over occasions is constant across people in $\mathcal{P}$. Because the error variables in latent variable models are not necessarily perfectly unreliable like the error variables in CTT, they can result from both sources of variability: the sampling of subjects and the sampling of occasions. That is, in MTT, the error variable can comprise a combination of both reliable and unreliable variables.

3 A Causal Theory of Error

In contrast to CTT, in which the error variable comprises all that is unreliable, the error variable in the latent variable model as defined in the previous section, comprises all predictors that are *unique* with respect to the other variables in $\mathbf{X}$. These unique predictors of the item score can be

perfectly reliable, perfectly unreliable or anything in between. We consider the error variables in latent variable models to reflect all unique causes of the item score that, together with Θ, fully determine the response variable. We interpret Θ as the common cause of the response variables (see e.g., Borsboom, Mellenbergh, & Van Heerden, 2003).

For the definition of a causal effect we follow the work of Pearl (2000). We let $do(Z = z)$ denote a surgical intervention on Z that sets Z to a particular value z without affecting the values of other variables and call any variable Z for which $P(X_j|do(Z = z)) \neq P(X_j)$ a cause of X_j. For example, a noisy surrounding that influences people's performance on a test implies that if one intervenes on this factor (e.g., by administering the test in a quiet place), the probability distributions of the item scores change. Similarly, if familiarity with certain words that are used in the description of a math item affects people's responses to that item, this implies that an intervention (e.g., teaching the meaning of these words before taking the test) changes the distribution of the responses on that item. Some of these causes characterize the person and others are part of the circumstances in which the items are administered. While such causal relations cannot be inferred from statistical associations alone, the equations in latent variable models can be interpreted as expressing *causal assumptions* (Pearl, 2009, 2012). For example, in the structural equation $Y = \mathbf{X}'\boldsymbol{\beta} + E$, $\mathbf{X}$ reflects a vector of hypothesized causes of Y and the error variable reflects all other causes of Y that account for the difference between $\mathbf{X}'\boldsymbol{\beta}$ and the actual values of Y (Chen & Pearl, 2013, p. 1).

Causal assumptions about error differ from statistical assumptions about error. Statistical assumptions concern the probability distributions of the variables, while causal assumptions pertain to the data-generating mechanism that gave rise to these probability distributions. As we will show later in this paper, many model assumptions that define latent variable models are assumptions about error, from which we conclude that an important part of reasoning about psychometric models is reasoning about errors. While these assumptions about error are statistical, a causal theory of error helps us reason about these assumptions by providing an interpretation in terms of the data-generating mechanism. In the remainder of this article, we use 'the causal theory of error'[3] to refer to the theory that we discuss here but note that alternative causal theories of error could be formulated.

Definitions of the variables. We distinguish two types of causes of the response variable that together constitute the error variables in MTT. *Circumstance variables* are those causes that are perfectly unreliable, and *characteristic variables* are those causes that are perfectly reliable. Because the models we consider here assume local independence, this implies that the error variables only contain variance that results from causes that are unique to a single item. The MTT models considered in this paper thus assume that all characteristic and circumstance variables are *unique* causes of the response variable. In the following, *characteristic variables* and *circumstance variables* are therefore used to refer to causes unique to a single item.

Let Ψ_j be the composite of all G circumstance variables unique to item j, $\Psi_j = \delta_{1j}\Psi_{1j} + \delta_{2j}\Psi_{2j} + ... + \delta_{Gj}\Psi_{Gj}$, weighted by their causal effect on X_j (δ_{1j} to δ_{Gj}). Ψ_j (as well as any constituent variable Ψ_{gj}) is a random variable of which the sample space is defined by randomly selecting pairs of subjects and occasions from $\mathcal{P} \times \mathcal{K}$. For a specific person i, Ψ_{ij} is a random variable of which the probability space is defined by randomly selecting measurement occasions from $\mathcal{K}$. Let Φ_j refer to the composite of all H characteristic variables unique to item j, $\Phi_j = \gamma_{1j}\Phi_{1j} + \gamma_{2j}\Phi_{2j} + ... + \gamma_{Hj}\Phi_{Hj}$, weighted by their causal effect on X_j (γ_{1j} to γ_{Hj}). Both Φ_j (as well as any constituent variable Φ_{hj}) and Θ are random variables over the domain $\mathcal{P} \times \mathcal{K}$. But, because Θ and Φ_j are perfectly reliable, they are constant across occasions in $\mathcal{K}$ and so their only source of probability is the sampling of subjects from $\mathcal{P}$. For a specific person i, θ_i and ϕ_{ij} are constants. The unique influences and Θ together determine[4] the response variable for item j, X_j,

$$X_j = \kappa_j + \lambda_j\Theta + \delta_j\Psi_j + \gamma_j\Phi_j, \tag{7}$$

where κ_j is an intercept. The error variable E_j is a mean-centered composite of all unique influ-

[3] In our causal theory of error, the error is interpreted as a composite of causes and as such one could also speak of a theory of 'causal error', in which error refers to the unique causes of an item, in contrast to, for example, a theory of 'errors as residuals', in which the error refers to the part of the item scores that is left unexplained.

[4] Since Equation (7) is a linear model, it would be sufficient to count something as a cause only if it affects the expected value, while in Pearl's account something is also a cause when it affects the variance but not the expected value. For example, Rubin's account of causality differs from Pearl's account because it excludes variables that only affect the variance (Markus, 2021), but would also be consistent with our causal theory of error.

ences,

$$E_j = C_j - \mathrm{E}(C_j), \tag{8}$$

where

$$C_j = \delta_j \Psi_j + \gamma_j \Phi_j. \tag{9}$$

The expected value of C_j together with κ_j constitute the intercept ν_j in the factor model,

$$\nu_j = \kappa_j + \delta_j \mathrm{E}(\Psi_j) + \gamma_j \mathrm{E}(\Phi_j). \tag{10}$$

$\mathrm{E}(\Psi_j)$ is the expected value of the circumstance variable across pairs of people and circumstances $\mathcal{P} \times \mathcal{K}$ and $\mathrm{E}(\Phi_j)$ is the expected value of the characteristic variable across people in $\mathcal{P}$. The inclusion of the expected value of the characteristic variables in the intercept means that the intercept (or difficulty of the item) can differ over groups of people with different characteristics. For example, a math item that is written in English is more difficult for people who have difficulty reading English. In fact, the presence of characteristic variables implies that it is possible to create groups that differ on their expected value for Φ_j, which implies item bias. We discuss item bias in more detail in a later section. The inclusion of the expected value of the circumstance variables in the intercept means that the intercept (or difficulty) can differ over different circumstances. For example, a math item is more difficult if the item is administered only in situations in which people are tired, while other items are not. κ_j represents the part of the intercept that does not depend on the expected value of the circumstance variable or characteristic variable. This means that also in the absence of item-specific causes, items can differ in their difficulties. The lack of a subscript i in the parameters λ_j, δ_j and γ_j indicates the assumption that the strength of the causal effects is constant across people. Relaxing this assumption results in a more complex model.

We also assume that the effects of all item-specific causes are captured in a single variable, E_j. In MTT, the error variable is assumed to have a normal or logistic distribution (the latter is simpler to work with and closely approximates the shape of a normal distribution; Lord & Novick, 1968, section 16.5). The idea that the distribution of error approximates a normal distribution is theoretically justified by assuming that the error is the sum of many independent errors and by the central limit theorem, which states that the sum of many independent variables that need not be normally distributed, will approximate a normal distribution (Durrett, 2019; Stigler, 1986). In the causal interpretation of error, the error is the sum of many independent causes.[5] We assume that the characteristic and circumstance variables are also additive, so that their sum captures all the information relevant to the item response. Another way of looking at this is that we assume that conditioning on C_j makes the set of all characteristic and circumstance variables independent of the item score. This is similar to the role of propensity scores that are used in studies about the effects of treatments, in which one wants to match treatment and control pairs on a large set of covariates. The set of covariates can be collapsed into a single variable, the propensity score, for which it is the case that the conditional distribution of the covariates given the propensity score is the same for people in the treatment and control group (Rosenbaum & Rubin, 1983). The assumption that the sum C_j captures all relevant information of the characteristic and circumstance variables, may be considered implausible, in which case one could choose to relax it and build a more complex model that, for example, allows for interactions between the causes. Importantly, the implausibility of such assumptions can be reason to change the model, but does not undermine the causal interpretation of these assumptions.

The switch from error as a residual to error as a composite of causes of the response variable has as a consequence that the assumption of error being uncorrelated with Θ is harder to justify (Markus, 2010). After all, the assumption can no longer be justified from the assumption that error only consists of random noise that is uncorrelated to anything. Especially for the characteristic variables it is difficult to justify that they are uncorrelated to Θ. For example, suppose that only item j in a math test uses difficult language and therefore is also influenced by people's language ability. A correlation between math ability (Θ) and language ability (characteristic variable of item j) would then violate this assumption. At the same time, if there is systematic error, the presence of which can be empirically tested, this points at the presence of characteristic variables, but also immediately violates the justification that the error is fully random and uncorrelated to anything.

[5] For simplicity, we here make the strict assumption of independence, but it is also possible to assume weaker forms of this assumptions for which it is still the case that the sum approximates a normal distribution (McLeish, 1974; White, 2001).

As such, a causal interpretation of error highlights how the assumption of error being uncorrelated to Θ is not trivial (Bollen & Pearl, 2013; Hayduk & Pazderka-Robinson, 2007; Markus, 2010).

The parameters δ_j and γ_j in the model in Equation (7) are not uniquely identified, and repeated measures would be needed to separate variance due to characteristic and circumstance variables. However, our goal here is to use the causal model in Equation (7) to interpret model assumptions as well as parameters in well-known latent variable models. For example, in IRT and CFA, the error variables all have mean zero, but items can have different intercepts. The causal theory of error provides an interpretation of this item parameter; ν_j reflects the expected value of the sum of all unique causal effects on the responses to item j.

The causal theory of error rests on the assumption that the variables in the psychometric models that we discuss influence each other in a deterministic system. That is, although some consider the act of responding to an item itself a stochastic event (de Boeck & Wilson, 2004, pp. 49–50), for example, because the cognitive process resulting in the response is a stochastic process (Van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011), we assume that responding is not intrinsically stochastic but that stochasticity in the responses conditional on Θ results from sampling a person from the subset of $\mathcal{P}$ for whom $\Theta = \theta$, and sampling a measurement occasion from $\mathcal{K}$.[6]

The response variable X_j is thus decomposed into the sum of three influences: (1) A common factor Θ that is shared with the other response variables in $\mathbf{X}$, (2) reliable unique factors that compose Φ_j, and (3) unreliable unique factors that compose Ψ_j. Next, we discuss what these unique causes that compose Φ_j and Ψ_j could be.

3.1 Circumstance Variables

The circumstance variables that are included in the error variables of MTT models are perfectly unreliable causes of the scores on a unique item. These are variables that are not controlled in the measurement procedure so that they vary over replications of the measurement procedure and are sometimes referred to as *random disturbances* (Trendler, 2009). Variability in the item scores that results from circumstance variables is called *random error* or *measurement error*. Rozeboom (1966, p. 380) introduces the notion of *circumstances* to refer to variables in the measurement procedure that are causally relevant to the item score:

> "the score actually assigned to a subject on a given test occasion is determined not merely by the subject's own attributes, but by a great many details of the surrounding circumstances as well."

Rozeboom gives a hypothetical example of a body weighing procedure in which some variables are controlled (e.g., it is part of the measurement procedure that the subject wears certain clothing) and some variables are uncontrolled (e.g., the air temperature and the subject's foot position are not standardized over repeated measures). Because the air temperature and subject's foot position vary over measurement circumstances, they produce variations in the responses over repeated measurements.

The circumstances Rozeboom (1966) describes are what we call *occasions* throughout this paper. Ellis and Van den Wollenberg (1993) explain occasions as including all uncontrolled factors that determine the item-score pattern of a subject, such as uncontrolled environmental variables but also temporary changes in the person, such as unpredictable mood fluctuations and concentration lapses. Ellis and Van den Wollenberg base their interpretation of the variability over occasions on Lord and Novick (1968, p. 38) who describe the *error-score random variable* as "a disturbance that is due to a composite of a multitude of factors not controlled in the measurement procedure". This error variable consists of factors associated with the physical measurement process, the measurement environment and temporal fluctuations in the person (Lord & Novick, 1968, pp. 38–39).

Examples of circumstance variables may be temporary changes in the person such as tiredness or distraction in the context of an ability test, or mood fluctuations that influence how someone responds to a questionnaire about life satisfaction. Another example is that one may be more

[6] With the assumption of a *deterministic system* we mean that all causes of the item response together fully determine the item response. Readers who do not want to make this assumption in order to leave room for fundamental stochasticity will disagree that *all* variability in the error variables reflects unique causes. However, also if only some part of the variance that is flagged as *error* reflects unique causes of the item, assumptions about these causes have implications for the models and the ability to reason about these causes puts researchers in a better position to motivate modeling choices.

inclined to endorse the personality item "I am a very social person" after going out the night before and meeting new people than after a night going to bed early.

3.2 Characteristic Variables

We defined characteristic variables as variables that, just like Θ, vary *over* persons and are fixed *within* a person over repeated measures of the same item. In contrast to Θ, characteristic variables are causally relevant to the score on a *unique* item. These are what Ellis and Van den Wollenberg (1993) call *item specific traits*, what Shealy and Stout (1993) call *nuisance determinants* and what Trendler (2009) calls *systematic disturbances*. In the multitrait-multimethod literature these effects are referred to as *item-specific methods* (Thielemann, Sengewald, Kappler, & Steyer, 2017). Variability in the error variable that results from characteristic variables is called *systematic error*.

A culture-specific word in a math item can have as a result that 'knowledge of the meaning of that word' causes some subjects to give a correct response to this item, while others not familiar with the meaning of that word give an incorrect response to the item, even though they all have the same math ability level θ. This characteristic of knowing the word or not varies over people not over measurement occasions. That is, by repeatedly administering the math test, it will be the same people that do and do not know the word. In that case, all variability in the variable 'word knowledge' is true-score variance, that is, the variable is perfectly reliable. Another example of a characteristic variable is whether a person taking an intelligence test has seen a certain movie that gave information relevant to one of the items. The variable 'having seen movie x' may not be related to people's intelligence but can help a person who saw the movie to solve an item that someone with the same intelligence but who has not seen that movie cannot solve. This other person might know the correct answer to an item that draws on knowledge about biology, because their mother works in that discipline. These two persons with the same level of intelligence may have the same total score, while having different items that they would consistently answer correctly, when repeatedly presented with the same item.

In sum, the causal theory of error defines two sources of error variability in latent variable models. The unidimensional latent variable model defined in the previous section allows for both sources. Later in this article, we investigate assumptions about these two sources of error, such as the absence of circumstance variables or characteristic variables. However, note that we show how such assumptions about the sources of error follow from assumptions in the model and they do not reflect our theories about what we consider plausible. That is, while assumptions in the model might imply that error consists of only one of these two sources of error, it is an empirical question whether that assumption is true, and in many cases the researcher might consider it more plausible that the error consists of a mix of both sources.

The causal theory of error that we proposed here has elements in common with two existing psychometric theories: Latent State Trait theory, and generalizability theory. In the next section, we explicate the connections between these approaches.

4 Relating the Causal Theory of Error to Other Theories

The causal theory of error is a theory about error that we formulated using the MTT (CFA and IRT) framework. In this section, we discuss how the causal theory of error relates to other psychometric theories. Earlier, we explained that the causal theory of error implies for Classical Test Theory that all characteristic variables are part of the true score variable, while the error variable in CTT consists of only circumstance variables. Now we turn to Latent State Trait theory and generalizability theory.

4.1 Latent State Trait Theory

The way that the error variable is composed of Ψ_j and Φ_j is in some ways similar to how in Latent State Trait (LST) theory (Steyer, Schmitt, & Eid, 1999) the latent common factor η is composed of a latent trait variable ξ that is stable over situations[7], and a state residual variable

[7]In the sense that trait variables are stable attributes of a person that are not influenced by the situation, the characteristic variable Φ_j and latent variable Θ can be considered trait variables. After all, Φ_j and Θ were defined as constant across occasions in $\mathcal{K}$, having the sampling of subjects from $\mathcal{P}$ as their only source of probability. However, the causal theory of error does not rely on the common factor Θ to be a trait variable, and the models discussed in

ζ that varies over situations. Similar to LST theory, the causal theory or error explains part of the variability in item scores by situational factors, such as whether a person slept a few hours versus a full night, before entering the measurement procedure (Steyer et al., 1999). However, in the causal theory of error, these factors explain the variability in people's observed scores across replications (i.e., what Lord & Novick, 1968, call the person's *propensity distribution*[8]). That is, people do not repeatedly obtain an observed score equal to their true-score *because* of such factors. In LST, however, the true-score is defined conditional on the person in the situation, so that people have a propensity distribution for each situation. As such, an individual's variability over different situations is variability in true-scores, not in error scores. In the causal theory of error,

$$\begin{aligned} X_j &= \kappa_j + \lambda_j \Theta + \delta_j \Psi_j + \gamma_j \Phi_j \\ E_j &= \delta_j \Psi_j + \gamma_j \Phi_j - \mathrm{E}(\delta_j \Psi_j + \gamma_j \Phi_j) \\ \nu_j &= \kappa_j + \mathrm{E}(\delta_j \Psi_j + \gamma_j \Phi_j) \end{aligned}$$

whereas in LST,

$$\begin{aligned} X_{jk} &= \xi_{jk} + \zeta_{jk} + E_{jk} \\ T_{jk} &= \xi_{jk} + \zeta_{jk}, \end{aligned}$$

where ξ_{jk} and ζ_{jk} are the latent trait variable and latent state residual for measurement j at the kth occasion of measurement, respectively. The error variable E_{jk} in LST is different from E_j in the causal theory of error, both because E_{jk} is conditional on a specific measurement occasion k while E_j is not, and also because E_{jk} in LST comprises only random measurement error, while E_j in the causal theory of error comprises all sources of variance in X_j that are unique to item j, including both random error and systematic error. To show the similarities and differences between LST theory and the causal error theory proposed here, we structured Table 1 in a similar way as Steyer et al.'s (1999) table that formulates LST theory. For our purposes here, an important difference with LST is that in the causal theory of error, the error variables are causes rather than difference variables (i.e., the difference between X_{jk} and the expectation of X_{jk} conditional on the person in the situation; Steyer et al., 1999)[9], and as such the causal theory of error seeks to interpret error variability. In contrast to LST, the error variable comprises all that is unique, and therefore both unreliable and reliable item-specific causes can make up the error variable.

Although occasions are hypothetical and only one is realized at the moment of measurement, one could view the administering of repeated measures of item j as a way of realizing multiple occasions that can be used to separate Ψ_j and Φ_j, because Ψ_j is assumed to vary and Φ_j is assumed to be fixed across occasions. The repeated measures should be close enough in time that it is realistic to assume that the target attribute and characteristic variables have not changed over time (e.g., no developmental processes or experiences in life that change the values of characteristic variables). Several existing models in the LST framework can be used to separate circumstance variables from characteristic variables by estimating item-specific method effects (see e.g., Schmitt & Steyer, 1993; Thielemann et al., 2017). Just like characteristic variables, method effects can be interpreted as causal effects on the item score (Maul, 2013; Pohl, Steyer, & Kraus, 2008). The method effects that are unique to an item are the same as those we call *characteristic variables*. For estimation purposes, all these models assume that the method-effect variables are either shared among different items relying on the same method (e.g., multiple math items relying on written reports in contrast to items relying on oral reports; see Pohl et al., 2008), or among repeated administrations of the same item. In the latter case, longitudinal data can be used to identify method effects that are unique to an item (see e.g., Thielemann et al., 2017).

When estimating LST models, the measurement error variable will reflect circumstance variables and the item-specific method effect will reflect characteristic variables. As such, assumptions about the absence or presence of characteristic and circumstance variables can be directly tested. However, the causal theory of error most importantly serves as a theoretical framework that can be used to reason about model assumptions in existing psychometric theories.

this article can easily be extended to include state variables as common factors.

[8] Not to be confused with *propensity probability* (Popper, 1959).

[9] This is close to how measurement error is defined in CTT, except that it is conditional on the situation. In contrast, the causal theory of error is closer to Latent Variable Theory, in which the latent variable represents a psychological attribute, than it is to CTT, in which true-scores have a operationalist definition as the expectation of the observed score X_j over replications (Borsboom, 2005).

Table 1: Formulation of the causal error theory.

Over persons	For specific person i	
Exogenous variables		
Θ	θ_i (parameter)	latent trait variable
Φ_j	ϕ_{ij} (parameter)	characteristic variable
Ψ_j	Ψ_{ij} (variable)	circumstance variable
Endogenous variables		
$X_j = \kappa_j + \lambda_j\Theta + \gamma_j\Phi_j + \delta_j\Psi_j$	$X_{ij} = \kappa_j + \lambda_j\theta_i + \gamma_j\phi_{ij} + \delta_j\Psi_{ij}$	item response variable
$E_j = \gamma_j\Phi_j + \delta_j\Psi_j - \mathrm{E}(\gamma_j\Phi_j + \delta_j\Psi_j)$	$E_{ij} = \gamma_j\phi_{ij} + \delta_j\Psi_{ij}$	error variable
Expected values		
$\kappa_j + \gamma_j\mathrm{E}(\Phi_j) + \delta_j\mathrm{E}(\Psi_j) = \nu_j$		intercept
$\mathrm{E}(E_j) = 0$		
Decomposition of variances		
$\mathrm{Var}(X_j) = \lambda_j^2\mathrm{Var}(\Theta) + \gamma_j^2\mathrm{Var}(\Phi_j) + \delta_j^2\mathrm{Var}(\Psi_j)$	$\mathrm{Var}(X_{ij}) = \mathrm{Var}(E_{ij}) = \delta_{ij}^2\mathrm{Var}(\Psi_{ij})$	
$\mathrm{Var}(E_j) = \gamma_j^2\mathrm{Var}(\Phi_j) + \delta_j^2\mathrm{Var}(\Psi_j)$		
Covariances		
$\mathrm{Cov}(\Theta_j, E_j) = \mathrm{Cov}(\Theta_j, \Phi_j) =$ $\mathrm{Cov}(\Theta_j, \Psi_j) = \mathrm{Cov}(\Phi_j, \Psi_j) = 0$		

Note: if there are no circumstance variables (i.e., $\mathrm{Var}(\Psi_{ij}) = 0$), then the item response variable and error variable for a specific person i are parameters (x_{ij} and ε_{ij} respectively).

4.2 Generalizability Theory

Generalizability theory is a test theory that differentiates between different sources of error (Briggs & Wilson, 2007; Shavelson & Webb, 1981), making it relevant to consider how generalizability theory links to the causal theory of error. In generalizability theory, a measurement is a sample from a universe of admissible observations, and is characterized by conditions such as the occasion of measurement, the particular items used, and the particular rater or tester (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1981). The universe of admissible observations consists of the observations under all different conditions that yield an acceptable basis for decision-making. The best basis for a decision would be a person's mean score over all acceptable observations (called the *universe score*), and it thus becomes an important question how well the sample *generalizes* to the universe (Cronbach et al., 1972).

Generalizability theory decomposes the variance of observed scores into the contributions of different facets, where a facet is a set of conditions of a certain type under which scores can be observed. For example, in a set up in which people's scores on 10 questions are observed, there are 10 conditions of the type 'question', which is one facet over which the scores vary. If there are multiple raters who score people on the multiple items, then there is an additional facet 'raters'. By adding more facets to the design, more sources of variation can be disentangled from the residual. In contrast to CTT, generalizability theory thus provides a framework to handle other sources of variation than just random error. In a one-facet crossed design, each person has scores on the same set of items, in which case the response variable X_{ij} can be decomposed as

$$\begin{aligned} X_{ij} &= \mu && \text{(grand mean)} \\ &\quad + \mu_i - \mu && \text{(person effect)} \\ &\quad + \mu_j - \mu && \text{(item effect)} \\ &\quad + X_{ij} - \mu_i - \mu_j + \mu && \text{(residual)}, \end{aligned} \tag{11}$$

where μ represents the grand mean across persons and admissible items, μ_i represents the person-specific mean across admissible items, and μ_j represents the item-specific mean across persons. In this context, *error* is defined as the residual that is left after subtracting item and person effects from the grand mean (Briggs & Wilson, 2007). This error component contains person by item ($i \times j$) effects as well as other sources of error that result in a deviation between X_{ij} and the sum of the first three components in Equation (11).

The grand mean is fixed while the item effect, person effect and residual are random variables. The person effect $\mu_i - \mu$, which is distributed over people, is comparable to Θ in MTT. People with

a positive person effect score above average on the total of admissible items, while people with a negative person effect score below average. Differences in ability can therefore be represented in different person effects (Eggen & Sanders, 1993). The difficulty of an item is captured in the item-specific mean, μ_j, which is a variable that is distributed across items. An item with a positive item effect $\mu_j - \mu$ is easier than an item with a negative item effect. While in this set up both items and persons are random, it is common in MTT to consider only persons as random and items as fixed[10]. In that case, μ_j is an item-specific fixed parameter, similar to the difficulty parameter in IRT and the intercept in factor modeling.

To relate generalizability theory to the causal theory of error, we can connect the characteristic and circumstance variables to the different variance components of the residual. Characteristic variables vary over people and can have different means across items. In the causal theory of error, the item-specific mean across people, $\mathrm{E}(\Phi_j)$, is part of the intercept: $\nu_j = \kappa_j + \delta_j \mathrm{E}(\Psi_j) + \gamma_j \mathrm{E}(\Phi_j)$ (see Equation (10)). In generalizability theory, $\mathrm{E}(\Phi_j)$ is part of μ_j; a higher mean of the characteristic variable makes the item easier. The mean-centered characteristic variable, $\Phi_j - \mathrm{E}(\Phi_j)$, is captured in the $i \times j$ interaction effect that is part of the residual in the above generalizability set-up. For fixed items, the $i \times j$ effect is a variable that varies over persons. The characteristic variable can indeed be interpreted as an interaction between people and item difficulty; an item is more difficult for people with a lower value on the characteristic variable. The circumstance variable varies over occasions. In the above set-up, occasions are not included as a separate facet. For that reason, the circumstance variable is one of the sources of error that is part of the residual.

If the items are not fixed but random, characteristic and circumstance variables are also the result of another source of variability, which is the sampling of items. To relate this source of variation to the examples in the introduction, this would correspond to the counterfactual "if I had answered a different item, I would have scored differently." In that case, the variation due to the person $\times$ item interaction effect is no longer true score variance; with every repeated administration a new item is sampled, corresponding to sampling a new value on the characteristic variable in the causal theory of error.

The random sampling of items is also at the heart of the domain-sampling theory, which was developed as an alternative to CTT (Nunnally & Bernstein, 1994). In contrast to the true score defined as the expected value over repeated measures, domain-sampling theory defines the true score as the expected value over repeated samples of an item from a hypothetical domain of items. The logic of this theory formed the basis for generalizability theory (Furr & Bacharach, 2008). The sampling of items thus provides another sampling theory foundation for probability in item scores that applies to contexts where the items are considered random (Holland, 1990).

In the next sections, we demonstrate how the causal theory of error can help interpret assumptions in MTT. More specifically, in the following sections we discuss (1) how common latent variable models are defined by assumptions about error, (2) how the two different chance experiments that Holland (1990) puts forward as possible sampling theory foundations of IRT models imply assumptions about error, and (3) how item bias and local homogeneity are assumptions about error.

5 Error Assumptions of Latent Variable Models

In this section, we show that different assumptions about the error variables result in different latent variable models. We summarize this overview in Table 2. The latent variable models we consider all entail the following three assumptions: (1) Θ is unidimensional, (2) $\mathbf{X}$ is conditionally independent given Θ, and (3) $P(X_j > x \mid \Theta)$ is monotonically nondecreasing in Θ (Junker & Ellis, 1997).

An example of a model that entails these assumptions is the unidimensional CFA model, $X_j = \nu_j + \lambda_j \Theta + E_j$ (Equation (6)), with items scored such that the factor loadings are nonnegative and with the error variables being independent of each other and of Θ (Junker & Ellis, 1997). In the causal theory of error, this CFA model allows for both characteristic and circumstance variables

$$X_j = \kappa_j + \lambda_j \Theta + \gamma_j \Phi_j + \delta_j \Psi_j, \tag{12}$$

where $\kappa_j + \mathrm{E}(\gamma_j \Phi_j + \delta_j \Psi_j)$ is the intercept ν_j. Note that the difference between characteristic and

[10] It is also possible to assume that items are random in the IRT context, which brings the IRT model even closer to generalizability theory (see Briggs & Wilson, 2007, for the explicit link between generalizability theory and random effects IRT models).

circumstance variables is like that between specific factors and error factors that together make up the unique variance in factor analysis (Thurstone, 1935). However, a difference is that we give both factors a causal interpretation, so that the specific factor comprises all causes that vary across people but not across occasions, and the error factor comprises all causes that vary across occasions but do not differ in expected value across people. If, in addition, the assumption is included that the error variable is perfectly unreliable, Equation (6) represents Jöreskog's model of congeneric item responses (Jöreskog, 1971). In the causal theory of error, this assumption implies the absence of characteristic variables,

$$X_j = \kappa_j + \lambda_j \Theta + \delta_j \Psi_j. \tag{13}$$

The CFA modeling framework is closely related to the IRT framework in which a logistic or cumulative normal function connects binary or polytomous response variables to the latent variable. The CFA model can be described as an IRT model for continuous item responses (Mellenbergh, 1994a, 1994b) and the IRT model is equivalent to nonlinear factor analysis with binary items (Takane & De Leeuw, 1987). For simplicity we consider binary response variables.

In IRT, it is common to represent the *probability of the item score* as a function of Θ rather than presenting the *item score itself* as a function of Θ. For example, in the 2-parameter logistic (2-PL) model (Birnbaum, 1968)[11], the probability of a positive response to item j given θ, π_j, is defined by means of the following nonlinear function,

$$\pi_j = P(X_j = 1 \mid \theta) = \frac{\exp\Big(a_j(\theta - b_j)\Big)}{1 + \exp\Big(a_j(\theta - b_j)\Big)}, \tag{14}$$

where b_j is the difficulty parameter of item j and a_j is the discrimination parameter. This typical formulation of the IRT model does not make the error variable E_j explicit, and so, to apply the causal theory of error to IRT models, we consider an alternative formulation for the IRT model. In the latent response variable formulation, a latent variable $X_j^* = \Theta - b_j + E_j$ is thresholded to obtain the observed variable X_j (Muthén, 1978; Takane & De Leeuw, 1987; Tutz, 1990). For example, the latent response variable formulation of the 2-PL model is

$$X_j = \begin{cases} 1 & \text{if } \Theta - b_j + E_j > 0 \\ 0 & \text{if } \Theta - b_j + E_j \leq 0, \end{cases} \tag{15}$$

where E_j follows a standard logistic distribution. The discriminating power of item j, a_j, is a function of the variance of E_j. The error variable E_j is distributed according to a logistic distribution with mean zero and a scale parameter s_{E_j}, which is proportional to the standard deviation of E_j, σ_{E_j}. A smaller variance of E_j corresponds to a stronger discriminating power of item j, which is reflected in a steeper item characteristic curve at the inflexion point of the IRF. If the assumption is added that the variance of E_j is the same for all items, this model represents the 1-PL model, also known as the Rasch model (Fischer, 1995; Rasch, 1960).

If the assumption is added that the error variable is perfectly unreliable, then for a specific person i, E_{ij} has a probability distribution that is defined over measurement occasions,

$$X_{ij} = \begin{cases} 1 & \text{if } \theta_i - b_j + E_{ij} > 0 \\ 0 & \text{if } \theta_i - b_j + E_{ij} \leq 0. \end{cases} \tag{16}$$

This assumption of perfectly unreliable error with the additional assumption that each person with the same level θ has the same error variance, $\sigma^2_{E_{ij}}$, implies local homogeneity (Ellis & Van den Wollenberg, 1993), which is the condition that people with the same level θ have the same response probabilities. When the assumption of local homogeneity is added to the IRT model, the resulting model is called the homogeneous IRT model. The fact that local homogeneity can be expressed as an assumption about error will be important when we relate the causal theory of error to the notion of item bias and local homogeneity.

If the error variable is assumed to be normally distributed, Equation (15) represents the normal-ogive model (Lord, 1952; Lord & Novick, 1968, p. 365) in which the standard deviation of E_j, σ_{E_j}, functions as the discrimination parameter (Tucker, 1946). The latent response variable formulation

[11] Birnbaum (1968) included a unit scaling factor of 1.7 (i.e., $1.7\alpha_j$ instead of α_j in Equation 14) to achieve greater agreement between the normal and logistic models.

Table 2: Summary of how models differ based on diverging assumptions about errors.

Assumptions about error	Latent response formulation for X_j	Conditional probability formulation for X_j	Latent response formulation for X_{ij}	Conditional probability formulation for X_{ij}	Model
1) $E_j \sim N(0, \boldsymbol{\sigma^2})$ 2) E_j is perfectly reliable	$X_j = 1 \quad$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \Phi(\theta - b_j)/\boldsymbol{\sigma}$	$x_{ij} = 1 \quad$ iff $\theta_i - b_j + \varepsilon_{ij} > 0$	The random variable X_{ij} does not exist, only x_{ij}.	1-parameter Normal ogive model
1) $E_j \sim N(0, \boldsymbol{\sigma_j^2})$ 2) E_j is perfectly reliable	$X_j = 1 \quad$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \Phi(\theta - b_j)/\boldsymbol{\sigma_j}$	$x_{ij} = 1 \quad$ iff $\theta_i - b_j + \varepsilon_{ij} > 0$	The random variable X_{ij} does not exist, only x_{ij}.	2-parameter Normal ogive model
1) $E_j \sim \text{Logistic}(0, \boldsymbol{s})$ 2) E_j is perfectly reliable	$X_j = 1 \quad$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \frac{\exp(\boldsymbol{a}(\theta - b_j))}{\exp(1 + \boldsymbol{a}(\theta - b_j))}$	$x_{ij} = 1 \quad$ iff $\theta_i - b_j + \varepsilon_{ij} > 0$	The random variable X_{ij} does not exist, only x_{ij}.	1-PL model
1) $E_j \sim \text{Logistic}(0, \boldsymbol{s_j})$ 2) E_j is perfectly reliable	$X_j = 1 \quad$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \frac{\exp(\boldsymbol{a_j}(\theta - b_j))}{\exp(1 + \boldsymbol{a_j}(\theta - b_j))}$	$x_{ij} = 1 \quad$ iff $\theta_i - b_j + \varepsilon_{ij} > 0$	The random variable X_{ij} does not exist, only x_{ij}.	2-PL model
1) $E_j \sim N(0, \boldsymbol{\sigma^2})$ 2) E_j is perfectly unreliable 3) $\sigma^2_{ij} = \boldsymbol{\sigma^2}$ given θ	$X_j = 1 \quad$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \Phi(\theta - b_j)/\boldsymbol{\sigma}$	$X_{ij} = 1 \quad$ iff $\theta_i - b_j + E_{ij} > 0$	$\pi_{ij} = \Phi(\theta_i - b_j)/\boldsymbol{\sigma}$	homogeneous 1-parameter Normal ogive model
1) $E_j \sim N(0, \boldsymbol{\sigma_j^2})$ 2) E_j is perfectly unreliable 3) $\sigma^2_{ij} = \boldsymbol{\sigma_j^2}$ given θ	$X_j = 1 \quad$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \Phi(\theta - b_j)/\boldsymbol{\sigma_j}$	$X_{ij} = 1 \quad$ iff $\theta_i - b_j + E_{ij} > 0$	$\pi_{ij} = \Phi(\theta_i - b_j)/\boldsymbol{\sigma_j}$	homogeneous 2-parameter Normal ogive model
1) $E_j \sim \text{Logistic}(0, \boldsymbol{s})$ 2) E_j is perfectly unreliable 3) $\sigma^2_{ij} = \boldsymbol{\sigma^2}$ ($s_{ij} = \boldsymbol{s}$) given θ	$X_j = 1 \quad$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \frac{\exp(\boldsymbol{a}(\theta - b_j))}{\exp(1 + \boldsymbol{a}(\theta - b_j))}$	$X_{ij} = 1 \quad$ iff $\theta_i - b_j + E_{ij} > 0$	$\pi_{ij} = \frac{\exp(\boldsymbol{a}(\theta_i - b_j))}{\exp(1 + \boldsymbol{a}(\theta_i - b_j))}$	homogeneous 1-PL model
1) $E_j \sim \text{Logistic}(0, \boldsymbol{s_j})$ 2) E_j is perfectly unreliable 3) $\sigma^2_{ij} = \boldsymbol{\sigma_j^2}$ ($s_{ij} = \boldsymbol{s_j}$) given θ	$X_j = 1 \quad$ iff $\Theta - b_j + E_j > 0$	$\pi_j = \frac{\exp(\boldsymbol{a_j}(\theta - b_j))}{\exp(1 + \boldsymbol{a_j}(\theta - b_j))}$	$X_{ij} = 1 \quad$ iff $\theta_i - b_j + E_{ij} > 0$	$\pi_{ij} = \frac{\exp(\boldsymbol{a_j}(\theta_i - b_j))}{\exp(1 + \boldsymbol{a_j}(\theta_i - b_j))}$	homogeneous 2-PL model

Bold symbols denote differences between the 1-parameter models and 2-parameter models.
Note that the assumptions in the first column are not sufficient for the models in the sixth column, but only describe assumptions about error. In addition to the assumptions in the first column, the models in the sixth column (all examples of monotone unidimensional IRT models), also assume (1) unidimensionality, (2) local independence and (3) monotonicity (Ellis & Van den Wollenberg, 1993; Junker & Ellis, 1997).

of a normal-ogive IRT model is equivalent to the formulation that is used in factor analysis with dichotomous items (Christoffersson, 1975; Muthén, 1978; Takane & De Leeuw, 1987)

$$X_j = \begin{cases} 1 & \text{if } \lambda_j \Theta - t_j + E_j > 0 \\ 0 & \text{if } \lambda_j \Theta - t_j + E_j \leq 0. \end{cases} \tag{17}$$

where t_j denotes the threshold which is the negative of the intercept ν_j.

Altogether, Equation (15) represents the 2-PL model if E_j is distributed according to a logistic distribution. Similarly, Equation (15) represents the homogeneous 2-PL model if the assumption is added that for any specific person i the error score E_{ij} is distributed according to a logistic distribution with the same mean and variance for persons with the same level θ. Equation (15) represents the 1-PL model if, in addition, E_j has the same variance for all items. If E_j in Equation (15) is normally distributed, this represents the normal-ogive model. These results are summarized in Table 2.

In the causal theory of error, these model assumptions are assumptions about the characteristic and circumstance variables,

$$X_j = \begin{cases} 1 & \text{if } \kappa_j + \lambda_j \Theta + \delta_j \Psi_j + \gamma_j \Phi_j > 0 \\ 0 & \text{if } \kappa_j + \lambda_j \Theta + \delta_j \Psi_j + \gamma_j \Phi_j \leq 0. \end{cases} \tag{18}$$

For example, the 2-PL model allows that the variance of C_j, the sum of unique causes, can differ for different items whereas the 1-PL model assumes that these variances are equal. The assumption of local homogeneity implies the absence of characteristic variables, and that the variance due to circumstance variables is the same for people with the same level θ.

In the next four sections, we relate the causal theory of error to Holland's (1990) two rationales for stochasticity in response variables and to item bias, local homogeneity, and reliability.

6 Holland's Two Rationales

A statistical model is defined as a set of probability distributions on the sample space (McCullagh, 2002), where the sample space is the set of all possible outcomes of a chance experiment (Knight, 2000). The use of statistical models in analyzing item responses thus relies on the assumption that the response variables are *random variables*, that is, some chance experiment underlies their realizations. More specifically, in MTT the response variables have a probability distribution conditional on θ.

Holland (1990) compared the stochastic subject rationale and the random sampling rationale with respect to the randomness of response variables in IRT models. Consider an example in which X_j is a response variable with item responses to the item '$2 + 7 = \ldots$' and the targeted attribute is math ability. The random sampling rationale and stochastic subject rationale give a different interpretation to the phrase: "the conditional probability of a positive response to the item '$2 + 7 = \ldots$' given math ability level θ is .7". In the stochastic subject rationale, this phrase is interpreted as "each person with math ability level θ has a probability .7 of giving a correct response to the item '$2 + 7 = \ldots$' ". In the random sampling rationale, this phrase is interpreted as "the probability is .7 of sampling a person who gives a correct response to the item '$2 + 7 = \ldots$' from the population of people with math ability level θ". In this rationale, a fixed person's response to the item is *not* stochastic; the random sampling of people from a population is the only source of probability. The current section connects the causal interpretation of error to these chance experiments to show that the stochastic subject rationale implies that the error variable is perfectly unreliable while the random sampling rationale implies that the error variable is perfectly reliable.

In the stochastic subject rationale, $P(X_j = 1 \mid \theta) = .7$ means that for each person with attribute level θ the probability of responding positively to item j is .7. This implies that the probability of sampling a person who gives a positive response to item j from the population with attribute level θ, $P(X_j = 1 \mid \theta)$, equals the probability for any person i in that population to give a positive response, $P(X_{ij} = 1 \mid \theta)$. In the random sampling rationale, there is *no* stochasticity within persons (Holland, 1990, p. 579). $P(X_j = 1 \mid \theta) = .7$ means that in the population of people with attribute level θ, 70% of the people respond positively and 30% respond negatively. Because there

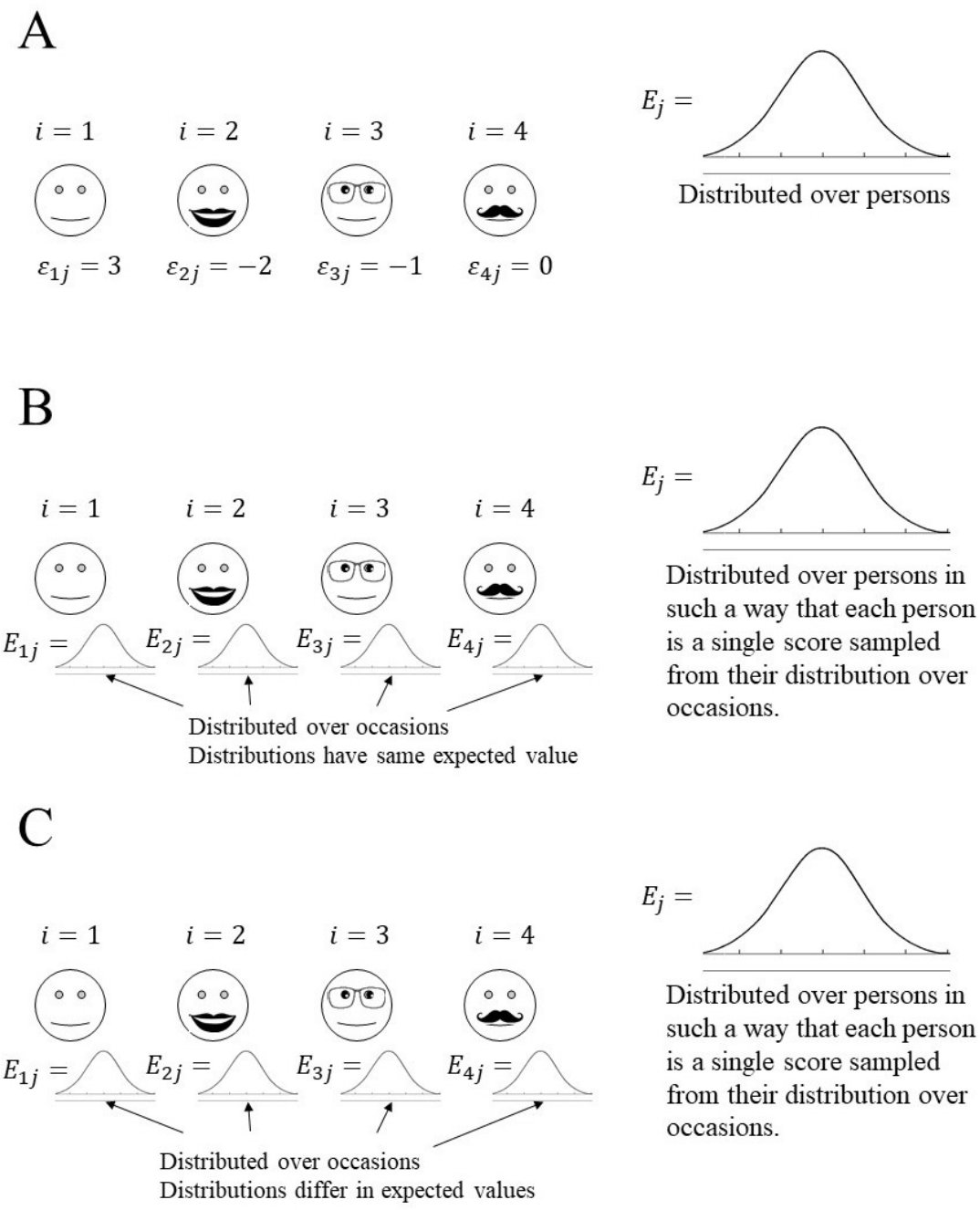

Figure 1: *Random Sampling and Stochastic Subjects as Different Sources of Randomness.*
Note. In panel A, the error variable E_j is perfectly reliable, which is consistent with the random sampling rationale; for a specific person i, ε_{ij} is a parameter. In panel B, the error variable E_j is perfectly unreliable, which is consistent with the stochastic subject rationale; for a specific person i, E_{ij} is a random variable defined over occasions, for which people have the same expected value, $\mathrm{E}(E_{1j}) = \cdots = \mathrm{E}(E_{4j})$. E_j in panel B thus does not reflect true-score variance. In panel C, the error variable E_j is partly reliable; for a specific person i, E_{ij} is a random variable, and people differ in their expected value for this variable.

is no stochasticity within persons, a person who gives a positive response does so with probability 1, as does a person who responds negatively.[12]

[12]By making the step that because there is no stochasticity within persons, people have either probability 1 or 0 of giving a positive response, we follow the work of e.g., Borsboom et al. (2003, p. 205) and Molenaar (1995, p. 6). Holland (1990), however, discussed an interpretation that hinges on defining a person as only existing at

Without error variables in the model, neither of the rationales above are sensible, because there is no variation in item scores between people with the same attribute level (random sampling rationale) nor is there variation in the item scores of a single person over repeated administrations of the item (stochastic subject rationale). Error thus is essential in explaining the different types of variation on which the two rationales are based. In the random sampling rationale, an interpretation of error is tantamount to explaining why people with the same attribute level θ nevertheless vary with respect to their item responses (i.e., some people with attribute level θ have probability 1 of responding positively, while others have probability 0). In contrast, in the stochastic subject rationale, an interpretation of error is tantamount to explaining why the responses of a person with a fixed attribute level θ vary over measurement occasions.

In the random sampling rationale, people have probability 1 or 0 of responding positively to an item, and hence the error variable is perfectly reliable. That is, E_j is constant over $\mathcal{K}$ and the only source of probability stems from the sampling of subjects from $\mathcal{P}$ (Panel A in Figure 1). In the causal theory of error, this corresponds to the hypothesis that all unique causes of the response variables are characteristic variables. These characteristic variables explain why people with the same attribute level vary in their item responses. In the stochastic subject rationale, people with the same attribute level have the same probability of responding positively to an item, which implies a perfectly unreliable error. In the causal theory of error, this implies the hypothesis that all unique causes of the response variables are circumstance variables. The presence of circumstance variables introduces a second source of probability, which is the sampling of measurement occasions from $\mathcal{K}$ (Panel B in Figure 1).

Perfectly reliable error is a necessary condition for the random sampling hypothesis, because in the random sampling hypothesis *all* variation in responses conditional on θ is true-score variance. After all, in the random sampling hypothesis, persons have response probabilities 1 or 0 and so for any person i the realized response x_{ij} equals the expected value $\mathrm{E}(X_{ij})$. Perfectly reliable error is also a sufficient condition for the random sampling hypothesis. If the variation in responses conditional on θ is perfectly reliable, then people have response probabilities 1 or 0.

Perfectly unreliable error is a necessary condition for the stochastic subject hypothesis, because in the stochastic subject hypothesis *no* variation in responses conditional on θ is true-score variance. In the stochastic subject hypothesis, people have the same response probabilities given θ and so their expected values $\mathrm{E}(X_{ij})$ (i.e., true-scores) are the same. However, perfectly unreliable error is not a sufficient condition for the stochastic subject hypothesis. Consider two persons 1 and 2, and let the sum of all characteristic and circumstance variables, C_j, be perfectly unreliable so that persons 1 and 2 have the same expected value for the error variable. Now suppose that person 1 has a smaller variance over repeated administrations of item j than person 2 has ($\sigma^2_{C_{1j}} < \sigma^2_{C_{2j}}$), then if $b_j < \theta$, person 2 has a higher probability of responding positively to item j than person 1, and if $b_j > \theta$, person 1 has a higher probability of responding positively to item j than person 2. For this reason, in addition to perfectly unreliable error, any pair of persons with the same attribute level also needs to have the same error variance over occasions ($\sigma^2_{C_{1j}} = \sigma^2_{C_{2j}}$ and thus $\sigma^2_{E_{1j}} = \sigma^2_{E_{2j}}$). Only if the distribution of C_{ij} is invariant over persons with the same level θ, the stochastic subject rationale holds and in this case, conditional on $\Theta = \theta$, C_j (and thus E_j) has as its only source of probability the sampling of measurement occasions from $\mathcal{K}$.

Holland's (1990) two rationales thus correspond to different hypotheses in the causal theory of error. Both hypotheses (the absence of characteristic variables and the absence of circumstance variables) are extremes of a spectrum and reality could be somewhere in between. If the error consists of both circumstance and characteristic variables, people will have response probabilities between 0 and 1 so that some of the stochasticity is within persons, but also people with the same level θ will differ in their response probabilities, so that the sampling of people is also a source of stochasticity (Panel C in Figure 1). In this mixed formulation, both random sampling and stochastic subjects contribute to the stochasticity (Ellis & Van den Wollenberg, 1993). In the next section, we relate the causal interpretation of error to the notion of item bias (Mellenbergh, 1989).

the moment that the item is administered. That is, in a repeated administration of an item, person i changes into person i^*, which means that inferences about person i do not generalize to later or earlier timepoints.

7 Item Bias

In this section, we show that the causal interpretation of error is also relevant for the interpretation of item bias. More specifically, we show that different forms of item bias correspond to different hypotheses about characteristic and circumstance variables.

An item j is unbiased with respect to a nominal variable G indicating group membership if and only if

$$F(X_j \mid \theta, G = g) = F(X_j \mid \theta), \tag{19}$$

where $F(.)$ denotes the (cumulative) distribution function (Mellenbergh, 1989). This definition of item bias is identical to the definition of measurement invariance (Meredith, 1993) and of subpopulation invariance (Ellis & Van den Wollenberg, 1993). It also implies Lord's (1980) definition of lack of bias (Meredith, 1993). In the special case of a dichotomous response variable X_j, Equation (19) implies that an item is unbiased with respect to G (Mellenbergh, 1989), if and only if

$$P(X_j = 1 \mid \theta, g) = P(X_j = 1 \mid \theta), \tag{20}$$

which means that the IRF of item j does not depend on group membership. Next, we relate the causal interpretation of error to item bias for both continuous response variables in the CFA framework and binary response variables in the IRT framework.

7.1 Item Bias for Continuous Response Variables

To see how the causal interpretation of error relates to the notion of item bias, consider Equations (7) to (10). For continuous response variables, Meredith (1993) showed that absence of bias requires that strict factorial invariance holds. A set of continuous response variables $\mathbf{X}$ is said to be strictly factorially invariant with respect to selection based on a grouping variable G if (1) the vector of factor loadings is equal over subpopulations based on G, (2) the vector of intercepts is equal over subpopulations based on G, and (3) the error variances are equal over subpopulations based on G (Meredith, 1993). In Equation (7), this means that λ_j is equal over groups, $\mathrm{E}(C_j) + \kappa_j = \nu_j$ is equal over groups, and $\sigma^2_{E_j} = \sigma^2_{C_j}$ is equal over groups.

We defined characteristic variables as unique reliable causal effects on the item response. The presence of characteristic variables Φ_j in the weighted sum C_j (i.e., at least one of the unique influences is a characteristic variable) implies that the population can be grouped in subpopulations based on Φ_j so that the expected value of the composite of causal effects varies across subpopulations; that is,

$$\text{presence of } \Phi_j \quad (\text{i.e.,} \quad \rho_{C_j C'_j} > 0) \implies \mathrm{E}(C_j) \neq \mathrm{E}(C_j \mid G_{\Phi_j} = g_{\Phi_j}), \tag{21}$$

where $\rho_{C_j C'_j}$ is the reliability of C_j, and G_{Φ_j} is a grouping variable based on Φ_j. The presence of a characteristic variable thus implies that the intercepts differ across groups based on G_{Φ_j}, and this violates strict invariance.[13]

The absence of characteristic variables (i.e., C_j is a weighted sum of only circumstance variables) is a necessary but not sufficient condition for lack of item bias; when C_j is perfectly unreliable, there may still be item bias because subpopulations differ with respect to the variance of C_j or with respect to the strength of the effect of Θ on the response variable (i.e., differences in λ_j).

7.2 Item Bias for Binary Response Variables

For binary response variables, we can use the latent variable formulation in which X^*_j is thresholded

$$X_j = \begin{cases} 1 & \text{if } X^*_j > 0 \\ 0 & \text{if } X^*_j \leq 0. \end{cases} \tag{22}$$

$$X^*_j = (\lambda_j \Theta - t_j) + E_j = \kappa_j + \lambda_j \Theta + C_j, \tag{23}$$

where the negation of the threshold $(-t_j)$ equals the intercept $\nu_j = \kappa_j + \mathrm{E}(C_j)$.

[13] When local independence is violated, it is also possible that a characteristic variable influences more than one item and as such induces item bias in multiple items.

The presence of a characteristic variable, Φ_j, implies that the population can be grouped in subpopulations so that the expected value of the composite of causal effects varies across subpopulations (i.e., $\mathrm{E}(C_j) \neq \mathrm{E}(C_j \mid g_{\Phi_j})$; see Equation (21)). As a result, it follows from Equation (23) that $\mathrm{E}(X_j^* \mid \theta) \neq \mathrm{E}(X_j^* \mid \theta, g_{\Phi_j})$, hence, $P(X_j = 1 \mid \theta) \neq P(X_j = 1 \mid \theta, g_{\Phi_j})$, which means that X_j is biased with respect to grouping variable G_{Φ_j}. This type of bias implies that threshold t_j in Equation (23) (and equivalently the intercept ν_j) differs over subpopulations based on G_{Φ_j}.

For binary response variables, an item is unbiased if and only if the IRF of X_j across all subpopulations coincides. That is, finding that either the discrimination parameter or the difficulty parameter differs over subpopulations indicates item bias. Because the difficulty parameter is a function of t_j and λ_j ($b_j = t_j/\lambda_j$; Muthén & Asparouhov, 2002), the bias that results from a characteristic variable manifests itself in different difficulty parameters b_j between subpopulations that are matched on the latent variable.

Uniform item bias exists when the IRFs for different groups do not coincide but when there is no interaction between Θ and group membership (Mellenbergh, 1982; Swaminathan & Rogers, 1990). That is, the difference between subgroups in the probabilities of a positive response is equal at all levels of Θ. For the 2-PL model, uniform item bias can be expressed in the hypothesis that the IRFs of the item are equally discriminating over groups and only differ with respect to the difficulty parameter. When the variance of C_j is equal over subpopulations (i.e., $\sigma^2_{C_j} = \sigma^2_{C_j|g}$), characteristic variables result in uniform item bias. After all, the discrimination parameter is a function of the variance of E_j ($a_j = \lambda_j/\sigma_{E_j}$; Muthén & Asparouhov, 2002), and $\sigma^2_{E_j} = \sigma^2_{C_j}$. The existence of different error variances over subpopulations (i.e., $\sigma^2_{C_j} \neq \sigma^2_{C_j|g}$) implies different discrimination parameters in these subpopulations and thus implies nonuniform bias.

We conclude that in the causal theory of error, if $\sigma^2_{C_j} \neq \sigma^2_{C_j|g}$, then item j is nonuniformly biased; that is, the discrimination parameter is a function of the variance of C_j. This conclusion is independent of whether the unique influences that compose C_j are characteristic variables or circumstance variables. If $\sigma^2_{C_j} = \sigma^2_{C_j|g}$, then there are two options: (1) if all unique influences are circumstance variables (i.e., error is perfectly unreliable), then there is no item bias; and (2) if at least one of the unique influences is a characteristic variable (i.e., error is at least partly reliable), then there is uniform item bias.

Although the presence of characteristic variables entails the presence of item bias in the technical sense formulated in Equation (19), in practice, item bias is typically investigated with respect to a small number of groups that map onto social categories. It is possible that each characteristic variable only has a negligible influence on the item responses and only a group variable based on a composite of many different characteristic variables would give groups (e.g., based on whether a person has read book x and whether the person has seen movie y and whether the person has had experience z, et cetera) with significantly different probabilities of a positive response given θ. Such group factors that do not map onto any relevant social categories are not typically what one is after in research on item bias. Therefore, if analyses show that a substantial part of the error variance is systematic, which indicates the presence of characteristic variables, the next step is to investigate whether this systematic part of the error variable correlates with meaningful group factors. For example, if it is more probable for people of a certain social category to have read book x, seen movie y and have had experience z (all contributing a little bit to a higher probability of a correct response), then one can detect item bias with respect to these social categories. But, it is also possible that the characteristic variables do not correlate with any social categories, in which case, in practice, such item bias will not be detected and is arguably also less relevant.

8 Local Homogeneity

The notion of item bias is closely related to the notion of local homogeneity. Local homogeneity holds when people with the same attribute level have identical response probabilities (Ellis & Van den Wollenberg, 1993). In contrast to the concept of item bias in which probabilities are defined for groups of people with the same attribute level θ, the assumption of local homogeneity takes the stochasticity explicitly to the subject level; each person with the same attribute level θ has the same probability of a positive response. Ellis and Van den Wollenberg (1993) showed that the stochastic subject interpretation holds if and only if the assumption of local homogeneity is met. Local homogeneity implies that there are no characteristic variables (Ellis & Van den Wollenberg, 1993, p.423). More generally, perfectly unreliable error is a necessary condition for local homogeneity.

Because we already showed that unreliable error is a necessary condition for the stochastic subject hypothesis but not a sufficient condition, it may come as no surprise that perfectly unreliable error is not a sufficient condition for local homogeneity either. Local homogeneity only holds if, in addition to perfectly unreliable error, given the same attribute level, each person has the same variance of C_{ij} over repeated administrations.

9 Coefficient α and test-retest correlation

Whereas the presence of characteristic variables implies item bias and violates the assumption of local homogeneity, the presence of circumstance variables has implications for the reliability of item scores. The following illustrates some of the implications of circumstance variables and characteristic variables for two measures of test-score reliability; coefficient α and the test-retest correlation. Equation (4) defined the reliability of an item in CTT. The reliability of a test can be defined in a similar way. Let X_+ denote the total score or test score

$$X_+ = \sum_{j=1}^{J} X_j. \tag{24}$$

The reliability of a test is defined as (Lord & Novick, 1968)

$$\rho_{X_+X'_+} = \frac{\sigma_T^2}{\sigma_{X_+}^2}, \tag{25}$$

where σ_T^2 is true-score variance of test X_+ and $\sigma_{X_+}^2$ is the total variance of test X_+.

The true-score variance, σ_T^2, is not available from data of a single test administration and thus alternative methods have been proposed to estimate the reliability from data of a single test administration (Sijtsma, 2009). One such coefficient is coefficient α (Cronbach, 1951; Guttman, 1945; Kuder & Richardson, 1937)

$$\alpha = \frac{J}{J-1}\Big(1 - \frac{\sum_{j=1}^{J}\sigma_{X_j}^2}{\sigma_{X_+}^2}\Big). \tag{26}$$

Coefficient α is shown to be equal to the reliability $\rho_{X_+X'_+}$ if the items in the test are essentially tau-equivalent (Novick & Lewis, 1967). Two variables are essentially tau-equivalent if and only if for two items j and k and a constant c_{jk} (Sijtsma, 2009)

$$T_j = T_k + c_{jk} \quad \text{for all items } j \neq k. \tag{27}$$

To understand the implications of characteristic variables and circumstance variables for test reliability, consider a model in which the response variables in $\mathbf{X}$ are caused by the same latent variable Θ, and each variable is also caused by characteristic variables Φ_1 to Φ_H and by circumstance variables Ψ_1 to Ψ_G. If the errors of the response variables in $\mathbf{X}$ are all composites of only circumstance variables (i.e., the variance of Φ_1 to Φ_H is 0), then all variance in X_j that is true-score variance is explained by Θ. That is, for $X_j = \nu_j + \lambda_j\Theta + E_j$ (see Equation (6)), the true-score variable, T_j, equals $\nu_j + \lambda_j\Theta$. Thus when λ_j (or equivalently the discrimination parameter in IRT) is the same over items, that is $\lambda_j = \lambda$, the items are essentially tau-equivalent and $\alpha = \rho_{X_+X'_+}$. In this case, both α and $\rho_{X_+X'_+}$ equal the proportion of variance in the test scores explained by Θ. When λ_j differs over items, $\alpha < \rho_{X_+X'_+}$ (α is a lower bound of the reliability; Guttman, 1945), and only the latter equals the proportion of variance in the test scores explained by Θ.

If the errors of the response variables in $\mathbf{X}$ are composites of both circumstance variables and characteristic variables, the items are not essentially tau-equivalent, in which case $\alpha < \rho_{X_+X'_+}$ and $\rho_{X_+X'_+}$ reflects the proportion of variance in the test scores explained by Θ and the characteristic variables. In the extreme, if the errors are composites of only characteristic variables (i.e., the variance of Ψ_1 to Ψ_G is 0), $\rho_{X_+X'_+} = 1$, whereas α is closer to 0 as a larger part of the variance of the test scores is caused by characteristic variables rather than Θ. When the variance of Θ is 0, that is, when the response variables are only caused by characteristic variables, $\alpha = 0$, while $\rho_{X_+X'_+} = 1$.

Another estimate of the reliability of a test is obtained by administering the test repeatedly and calculating the correlation between test scores. The test-retest correlation coefficient (Silk, 1977), which is the correlation between test scores on the same test administered at two measurement occasions, t_1 and t_2,

$$\rho_{\text{test-retest}} = \frac{cov(X_{+t_1}, X_{+t_2})}{\sqrt{\sigma^2_{X_+t_1}\sigma^2_{X_+t_2}}}, \tag{28}$$

quantifies the *relative consistency* of test scores, that is, "the consistency of the position or rank of individuals in the group relative to others" (Weir, 2005, p. 231). While α as a measure of reliability ignores true-score variance that is due to characteristic variables, the test-retest correlation coefficient does not. If the errors are a composite of only characteristic variables, $\rho_{\text{test-retest}} = 1$. If there is no common cause and the response variables are only caused by characteristic variables, then $\alpha = 0$, while $\rho_{\text{test-retest}} = 1$. If the errors are a composite of only circumstance variables, and the items are essentially tau-equivalent, then $\alpha = \rho_{\text{test-retest}}$. The more of the error variance is due to characteristic variables, the more the two indicators of reliability diverge. While the presence of characteristic variables results in α being an underestimate of $\rho_{X_+X'_+}$, characteristic variables do not result in a deviation between $\rho_{\text{test-retest}}$ and $\rho_{X_+X'_+}$. However, there can be other reasons why the test-retest correlation deviates from $\rho_{X_+X'_+}$ (e.g., systematic changes in the person over time, such as getting better at math over repeated measures of a math test, will result in $\rho_{\text{test-retest}} < \rho_{X_+X'_+}$).

Two additional reliability coefficients are McDonald's ω_t and ω_h (McDonald, 1999; Revelle & Zinbarg, 2009). These coefficients align with the factor-analyses approach to reliability, which is based on assumptions different from methods proposed in the context of classical test theory (e.g., Bentler, 2009; Raykov, 2001). In addition to approaches based on CTT and factor analyses, different reliability approaches based on generalizability theory and IRT exist (Sijtsma & Van der Ark, 2015), and because of the broadness of the topic, we limit the discussion to coefficient omega. Coefficient omega total (ω_t) is the sum of squared factor loadings on both general and group factors over the total test variance, while coefficient omega hierarchical (ω_h) is the sum of squared loadings on only the general factor over the total test variance. Just like α, both coefficients thus ignore true-score variance due to characteristic variables. For the unidimensional latent variable models considered in this article, the two omega coefficients are the same because these models include only a general factor, no group factors. In the presence of group factors, the coefficient omega hierarchical not only ignores true-score variance in the test scores that is due to characteristic variables but also that which is due to group factors, and thus treats variance due to group factors as error variance as well.

10 Practical Implications: Studying Error

In previous sections, we laid out implications of model assumptions for causes of the response variable, which can help substantive researchers choose what assumptions to include in a model based on what causal mechanisms they deem most plausible. As such, the causal theory of error is a tool to help reason about modeling assumptions in the MTT framework. For example, although assumptions such as absence of item bias and local homogeneity both imply that there cannot be any characteristic variables that cause the item scores, this might be considered implausible in practice. Rather than assuming the error variance to be fully unreliable or fully reliable, one can study empirically how much of the variance is due to characteristic variables (reliable variance) and how much is due to circumstance variables (unreliable variance).

Several existing models, both in the multitrait-multimethod literature and the LST literature, can be used to estimate item-specific method effects to separate variance due to characteristic variables from variance due to circumstance variables (see Geiser & Lockhart, 2012; Koch, Schultze, Holtmann, Geiser, & Eid, 2017; Pohl et al., 2008; Thielemann et al., 2017). These models are not only useful for indicating item bias due to characteristic variables, but also for studying substantive theories about error. For example, it might be hypothesized that in a memory task mostly circumstance variables influence the item scores while items in a knowledge test are more strongly influenced by characteristic variables. This hypothesis predicts that for items in a memory task, the variance of the measurement error will be larger than the variance of the item-specific method effects, while in a knowledge test this prediction is reversed.

Estimating the variance of the characteristic variables in these models relies on the use of repeated measures across which it can be assumed that the characteristic variables are constant. In some cases, however, it is plausible that characteristic variables do change over repeated measures, and one might want to study the change in characteristic variables over time. To do so, repeated measures within a time frame in which no change is assumed are required to statistically identify the item-specific effects, as well as measurements further apart in time that can be compared to identify possible change.

11 Hypotheses About the Unique Causes

In Table 3, we tie the previous sections together by summarizing how different hypotheses about the causes of the item scores connect to Holland's two rationales, the assumptions of local homogeneity and absence of item bias, reliability, and empirical predictions that can be used to test these hypotheses.

The first row in Table 3 represents the hypothesis that E_j consists of only circumstance variables and in addition the variance of these circumstance variables is the same across people. In this case, E_j is perfectly unreliable and the error score of item j for a person i, E_{ij}, is a random variable; the realization ε_{ij} is a single draw from person i's propensity distribution over possible circumstances for item j (panel B in Figure 1). This is what is implied by local homogeneity, absence of item bias, and the stochastic subject rationale. The empirical predictions of this hypothesis for repeated measures data are that there are no item-specific method effects, and coefficient α equals test-retest reliability.

The second row in Table 3 represents the hypothesis that E_j consists of only circumstance variables, but the variance of these variables over repeated measures differs across persons with the same level θ. Ellis and Van den Wollenberg (1993) mention "temporary changes within the person" (e.g., mood or energy fluctuations) as an example of something that is part of the occasion that determines the item response. However, some people may experience stronger change than other people, so that the variance of this variable differs over persons. If only the variance differs over persons (e.g., $\sigma^2_{C_{1j}} \neq \sigma^2_{C_{2j}}$ in Table 3), but the expected value is the same over persons ($\mathrm{E}(C_{1j}) = \mathrm{E}(C_{2j})$), the error variable is still perfectly unreliable. Unlike the hypothesis in the first row in Table 3, the hypothesis in the second row violates local homogeneity and absence of item bias. The hypothesis in the second row is also inconsistent with both the random sampling and stochastic subject rationales. Similar to the hypothesis in the first row, this hypothesis in the second row implies that there are no item-specific method effects, and coefficient α equals test-retest reliability.

A third and arguably more plausible hypothesis is that the error consists of both characteristic and circumstance variables, in which case C_{ij} varies over occasions, but has a different expected value for different people with the same level θ ($\mathrm{E}(C_{1j}) \neq \mathrm{E}(C_{2j})$ in Table 3). This situation is represented in the third row of Table 3. For example, responses to the Big Five Inventory (BFI) item "Is helpful and unselfish with others" (1 [strongly disagree] - 5 [strongly agree]; Benet-Martínez & John, 1998), which measures *Agreeableness*, will also be influenced by characteristics of the person that are specific to helpfulness and not shared with other agreeableness items (i.e., unique to that item). This specific helpfulness attribute can vary over testing occasions because some situations will make the attribute more salient (e.g., taking the item right after having done something helpful, or anything in the circumstance that reminds the person of helpful behavior). As such, the attribute is specific to the item, will fluctuate over occasions, and also has different expected values for different people. This hypothesis of partly reliable error violates the assumptions of local homogeneity and the absence of item bias, and is inconsistent with both the random sampling and stochastic subject rationale. Instead, this hypothesis implies that a combination of both stochasticity within subjects and the random sampling of people with different response probabilities are sources of stochasticity. That is, instead of 'only stochasticity within persons' or 'only the sampling of people and no stochasticity within persons', both stochastic mechanisms contribute, resulting in a mixed formulation (Ellis & Van den Wollenberg, 1993). For repeated measures data, this hypothesis predicts that there are item-specific method effects, but there is also measurement error (the item-specific variance that is not shared over repeated measures of the same item).

The fourth row of Table 3 represents the hypothesis that the error variable is a composite of only characteristic variables. In that case, E_j is perfectly reliable and the error score of item j for a

Table 3: Implications of Hypotheses About the Unique Causes of Item j for Two Persons, i = 1 and i = 2, Who Have the Same Attribute Level θ.

Hypotheses about C_j	Implications for conditional distributions and probabilities	Chance experiment	Empirical predictions
C_j is composite of only circumstance variables AND $\sigma^2_{C_{1j}} = \sigma^2_{C_{2j}}$	$\mathrm{E}(C_{1j}) = \mathrm{E}(C_{2j})$: **true** $\sigma^2_{E_{1j}} = \sigma^2_{E_{2j}}$: **true** $P(X_{1j} = 1) = P(X_{2j} = 1)$: **true** (= local homogeneity holds)	Stochastic subject hypothesis: **true** Random sampling hypothesis: **not true**	Item bias: **absent** Item-specific method effects: **absent** Measurement error: **present** Coefficient $\alpha = \rho_{\text{test-retest}}$
C_j is composite of only circumstance variables AND $\sigma^2_{C_{1j}} \neq \sigma^2_{C_{2j}}$	$\mathrm{E}(C_{1j}) = \mathrm{E}(C_{2j})$: **true** $\sigma^2_{E_{1j}} = \sigma^2_{E_{2j}}$: **not true** $P(X_{1j} = 1) = P(X_{2j} = 1)$: **not true**	Stochastic subject hypothesis: **not true** Random sampling hypothesis: **not true**	Item bias: **present** Item-specific method effects: **absent** Measurement error: **present** Coefficient $\alpha = \rho_{\text{test-retest}}$
C_j is composite of both circumstance and characteristic variables	$\mathrm{E}(C_{1j}) = \mathrm{E}(C_{2j})$: **not true** $\sigma^2_{E_{1j}} = \sigma^2_{E_{2j}}$: **true iff** $\sigma^2_{C_{1j}} = \sigma^2_{C_{2j}}$ $P(X_{1j} = 1) = P(X_{2j} = 1)$: **not true**	Stochastic subject hypothesis: **not true** Random sampling hypothesis: **not true**	Item bias: **present** Item-specific method effects: **present** Measurement error: **present** Coefficient $\alpha \neq \rho_{\text{test-retest}}$
C_j is composite of only characteristic variables	$\mathrm{E}(C_{1j}) = \mathrm{E}(C_{2j})$: **not true** $\sigma^2_{E_{1j}} = \sigma^2_{E_{2j}}$: **true** (for both individuals variance is 0) $P(X_{1j} = 1) = P(X_{2j} = 1)$: **not true**	Stochastic subject hypothesis: **not true** Random sampling hypothesis: **true**	Item bias: **present** Item-specific method effects: **present** Measurement error: **absent** Coefficient $\alpha \neq \rho_{\text{test-retest}}$ $\rho_{\text{test-retest}} = 1$

Note. The stochastic subject hypothesis means that each person with the same attribute level θ has the same probability of giving a positive response. The random sampling hypothesis means that each person has either probability 1 or probability 0 of giving a positive response.

person i, ε_{ij}, is fixed (panel A in Figure 1). As a result, of all people with the same level θ, depending on their fixed error score, any person has either probability 1 or 0 of responding positively to item j. So, for two people with the same level θ, one may have characteristics that result in a probability 1 of responding positively to item j while the other has characteristics that result in a probability 0. For some other item k with the same difficulty as item j, the second person may have other characteristics (and thus error score) that give them probability 1 of responding positively. More generally, for an infinite set of items, these two people would have the same proportion of items correct (Ellis & Junker, 1997; McDonald, 2003). Thus, in the long run, considering a sufficiently large sample of items, the characteristics that advantage and disadvantage the two people will counterbalance. This hypothesis also violates local homogeneity and the absence of item bias. The hypothesis is consistent with the random sampling rationale in which people have probability 1 or 0 for any response. For repeated measures, this hypothesis predicts that all variance that is specific to the item reflects item-specific method effects (only systematic error), and therefore there will be no measurement error variables when a method effect model is estimated from longitudinal data.

12 Discussion

While considerable thought goes into the interpretation of latent variables (e.g., this is the subject of construct validity research; Cook & Beckman, 2006; Cronbach & Meehl, 1955), the interpretation of the error variables is typically left implicit. Error variables are often called *residuals* (Kline, 2016, p. 13; Newsom, 2015, p. 2; Raykov and Marcoulides, 2000, p. 13), a remainder after the main part is accounted for. Yet, while it is difficult to define latent constructs such as intelligence, there are often strong ideas about what these constructs are *not*. Accidentally crossing the wrong answer, or getting distracted, do not count as low intelligence. Theories about what error is are thus also theories about what the construct is. We took error as a starting point and provided a theory of error in which error is the sum of myriad independent factors that influence the item score. We showed how different assumptions about error (1) imply different models, (2) have different implications for the chance experiment that makes the response variables random variables, and (3) have implications for item bias and local homogeneity. Table 2 provides a summary of (1) and Table 3 provides a summary of (2) and (3).

The causal theory of error scores that we laid out in this article should help researchers to motivate modeling choices based on the data-generating mechanism they consider plausible. For example, if researchers suspect influences on the item responses that are characteristic of the persons instead of of the measurement occasion, this should lead them to expect that there is item bias and that local homogeneity is violated. We therefore think it is also important to build substantive theories about the different sources of error in a specific context and to study error empirically. Because assumptions about error distinguish different psychometric models from each other, thinking more thoroughly about the causal nature of error puts researchers in the position to choose the right model.

The causal theory of error interprets the error variables in MTT. Because we reason from MTT, the errors have a specific structure that is constrained by assumptions in the MTT model. For simplicity, we considered a subset of MTT models, namely unidimensional latent variable models that are characterized by local independence, unidimensionality and monotonicity, and showed how the assumptions in those models translate to assumptions about the unique causes of the response variable. By adding or relaxing assumptions one arrives at different MTT models that allow for different causal structures. For example, relaxing the assumption of local independence will allow for causes that are shared between some but not all items, and adding the assumption of local homogeneity restricts the causal structure of error even more by stipulating that there can only be circumstance variables. Future research could study how the causal theory of error extends to other models than the ones we considered in this article.

Because we focused on models that are characterized by local independence, we restricted our use of *circumstance* and *characteristic variables* to reliable and unreliable causes that are *unique to a single item*, so that they are consistent with the assumptions in the model. However, many factors that we would consider *circumstances*, such as low energy levels because of a bad night's sleep, can affect more than one item. Circumstances that induce correlations by affecting more than one item are assumed to be absent by the assumptions of the unidimensional model characterized by local independence and unidimensionality. This does not mean that our theory is that such factors causing multiple items do not exist, but rather that the causal theory of error shows that

the unidimensional factor model is very strict.

If the assumptions of unidimensionality and local independence are relaxed, correlations between error variables can be interpreted as the presence of characteristic or circumstance variables that cause responses to multiple items (and thus are no longer unique to an item). However, in such cases it is also possible to explain the residual correlations by adding factors to the model, in which case such causes of multiple item responses would be included in factors common to subsets of items and would no longer be error. That is, what counts as *error* in the model will depend on modeling choices, so that the same cause of an item response variable can be considered *error* in the one model or modelled as a factor in a different model.

The many constraints on the causal structure of the error that we discussed in this article (e.g., absence of characteristic variables, characteristic and circumstance variables being unique to a single item and being uncorrelated with Θ) might give the impression that we consider these constraints plausible. However, the causal theory of error is not a theory about what causal structure we think underlies item responses, but a psychometric theory that maps assumptions about error in MTT models onto assumptions about the causes of item responses. As such, the causal theory of error may be used to help substantive researchers to think through the implications of model assumptions for the data-generating mechanism. If substantive theory about the causal structure is inconsistent with the causal structure that is implied by the model, one can opt for a more complex model that allows for structures that are consistent with the theory. In some cases, this might require the development of new models. In other cases, the researcher might decide to select a simpler model even though it includes assumptions that imply an implausible causal mechanism, or try to change the causal structure (e.g., by adjusting the items or circumstances in which the items are administered). In any case, the causal theory of error helps understand statistical assumptions about error and their implications, and, as we have shown in this paper, understanding assumptions about error is essential to understanding psychometric models.

A limitation of the presented theory is that only two sources of error are considered, and it would therefore be interesting to expand the theory to include sources of error that vary over other dimensions than people and circumstances. There is literature that deals with more than two dimensions that can be helpful in finding suggestions for potential dimensions to add to the causal theory of error (e.g., literature on generalizability theory, but also Reichardt, 2006, which concerns the different dimensions across which treatment effects can be estimated). Further research could then study what error assumptions in MTT models imply for such additional sources of error variation.

We have suggested the possibility of estimating item-specific method effects with LST models to estimate the variance in the error that is reliable. However, this solution relies on the assumption that the repeated measures approximate draws from someone's propensity distribution, and thus approximate the hypothetical experiment in which a person is brainwashed before answering the same item again (Lazarsfeld, 1959; quoted in Lord & Novick, 1968, p. 30). Changes in the characteristic variable over repeated measures (e.g., because of learning from the first time doing the item), or a tendency to respond consistently over repeated measures, threaten this assumption.

Consistent with Schmitt's (2006) and Pohl et al.'s (2008) interpretation of methods as sets of causes, we defined characteristic variables as causes. In contrast to their work in which a distinction is made between item-specific method effects and measurement error, we proposed an interpretation of the error variables in which the error variable as a whole is considered a set of item-specific effects, and instead distinguished between reliable and unreliable item-specific effects. The reliable item-specific effects (characteristic variables) are consistent with item-specific method effects while the unreliable item-specific effects (circumstance variables) provide a causal interpretation of measurement error.

We agree with Schmitt's (2006, p.24) view that "causes as components of a method do not differ from causes that appear in psychological theories". That is, we argue that characteristic variables are not inherently different from the attribute reflected in Θ; their difference is merely a result of which items are combined. The measured attribute is the common cause of all items, whereas characteristic variables are causes unique to a single item (or a few items, in case local independence is violated). In that sense, we agree with Maul (2013) that the term *method factor* is a bit of a misnomer because it misleadingly suggests that the factor represents a method rather than an attribute that is measured. We therefore prefer the term *method-specific trait*[14] (and *item-specific*

[14]Instead of *trait* one could also use the term *attribute,* but we here chose *trait* to distinguish it from state variables in LST theory, which are defined as attributes that vary over situations.

trait when unique to a single item; Ellis & Van den Wollenberg, 1993). The causal interpretation of error proposed in this article extends this view to include not only systematic error due to item-specific traits (characteristic variables) but also measurement error due to circumstance variables. Like methods, measurement error reflects item-specific causal effects, but, unlike methods, these causes vary over occasions instead of over persons.

Although the way we defined Θ and the characteristic variables, is consistent with how, for example, Ellis and Van den Wollenberg (1993) define a *latent trait* (i.e., as a variable that varies over persons but is constant across occasions), the causal theory of error does not require a trait perspective (see Mislevy, 2018, for different perspectives). Characteristic variables can be considered traits in the sense that they produce reliable variance in the item scores (which is how Ellis & Van den Wollenberg, 1993, use *trait*), but it is not necessary to view these variables as traits in the sense of context-invariant properties of a person. For example, the sociocognitive perspective, in contrast to a trait-perspective, views people's acting as mediated by socially constructed practices, such as shared language and culture (Mislevy, 2018). Characteristic variables can also refer to such social constructs, as long as they are attributes of people that are constant across occasions.

To summarize, the present article focused on the interpretation of error to interpret assumptions and fundamental concepts of modern test theory. The causal theory of error employed in this study has resulted in a framework that is useful in thinking through concepts such as stochastic response variables, item bias and local homogeneity. In addition, some well-known latent variable models were shown to follow from assumptions about error. We conclude that rather than approaching errors as residuals, the framework presented in this study shows that an understanding of error is essential for the interpretation of psychometric models.

References

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis (2nd ed.)*. London: Arnold. (Kendall's Library of Statistics 7).

Benet-Martínez, V., & John, O. P. (1998). Los cinco grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the big five in spanish and english. *Journal of Personality and Social Psychology*, *75*(3), 729–750. https://doi.org/10.1037/0022-3514.75.3.729

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*(1), 137–143. https://doi.org/10.1007/s11336-008-9100-1

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* (pp. 397–479). Reading, MA: Addison-Wesley.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*(3), 303–316. https://doi.org/10.1177/0049124189017003004

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305–314. https://doi.org/10.1037/0033-2909.110.2.305

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 301–328). Dordrecht, The Netherlands: Springer.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press. https://doi.org/10.1017/cbo9780511490026

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295x.110.2.203

Borsboom, D., & Molenaar, D. (2015). Psychometrics. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (pp. 418–422). Amsterdam, the Netherlands: Elsevier. https://doi.org/10.1016/b978-0-08-097086-8.43079-5

Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, *44*(2), 131–155.

Chen, B., & Pearl, J. (2013). Regression and causation: A critical examination of six econometrics textbooks. *Real-World Economics Review*, *65*, 2–20. https://escholarship.org/uc/item/1gf2k2mt

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*(1), 5–32. https://doi.org/10.1007/bf02291477

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, *119*(2), 166.e7-166.e16. https://doi.org/10.1016/j.amjmed.2005.10.036

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika*, *12*(1), 1–16. https://doi.org/10.1007/bf02289289

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washingtron, DC: American Council on Education.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability scores and profiles*. New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

de Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer Science+Business Media, LLC. https://doi.org/10.1007/978-1-4757-3990-9

Durrett, R. (2019). *Probability: Theory and examples* (5th ed., Vol. 49). Cambridge, NY: Cambridge University Press.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989x.5.2.155

Eggen, T. J. H. M., & Sanders, P. F. (Eds.). (1993). *Psychometrie in de praktijk*. Cito Instituut voor Toetsontwikkeling. https://www.cito.nl/kennis-en-innovatie/onderzoek/psychometrie-in-de-praktijk-boek

Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, *62*(4), 495–523. https://doi.org/10.1007/bf02294640

Ellis, J. L., & Van den Wollenberg, A. L. (1993). Local homogeneity in latent trait models. a characterization of the homogeneous monotone irt model. *Psychometrika*, *58*(3), 417–429. https://doi.org/10.1007/bf02294649

Fischer, G. (1995). Derivations of the rasch model. In G. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15–38). New York, NY: Springer. https://doi.org/10.1007/978-1-4612-4230-7_2

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage Publications, Inc.

Geiser, C., & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state–trait analyses. *Psychological Methods*, *17*(2), 255–283. https://doi.org/10.1037/a0026977

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, *42*(2), 139–167. https://doi.org/10.1111/j.2044-8317.1989.tb00905.x

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. https://doi.org/10.1007/bf02288892

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stoufer, L. Guttman, E. A. Suchman, P. L. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Studies in social psychology in world war II: Vol. IV. measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.

Hayduk, L. A., & Pazderka-Robinson, H. (2007). Fighting to understand the world causally: Three battles connected to the causal implications of structural equation models. In W. Outhwaite & S. P. Turner (Eds.), *The sage handbook of social science methodology* (pp. 147–171). Thousand Oaks, CA: Sage.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*(4), 577–601. https://doi.org/10.1007/bf02294609

Irwing, P., Booth, T., & Hughes, D. (Eds.). (2018). *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*. New York, NY: Wiley. https://doi.org/10.1002/9781118489772

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2),

109–133. https://doi.org/10.1007/bf02291393
Junker, B. W., & Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *The Annals of Statistics*, *25*(3), 1327–1343. https://doi.org/10.1214/aos/1069362751
Kline, R. B. (2016). *Principles and practice of structural equation modeling (4th ed.)*. New York, NY: The Guilford Press.
Knight, K. (2000). *Mathematical statistics*. Boca Raton, FL: Chapman & Hall/CRC. https://doi.org/10.1201/9780367805319
Koch, T., Schultze, M., Holtmann, J., Geiser, C., & Eid, M. (2017). A multimethod latent state-trait model for structurally different and interchangeable methods. *Psychometrika*, *82*(1), 17–47. https://doi.org/10.1007/s11336-016-9541-x
Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160. https://doi.org/10.1007/BF02288391
Lord, F. M. (1952). *A theory of test scores*. New York, NY: Psychometric Society.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum. https://doi.org/10.4324/9780203056615
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
Markus, K. A. (2010). Structural equations and causal explanations: Some challenges for causal sem. *Structural Equation Modeling*, *17*(4), 654–676. https://doi.org/10.1080/10705511.2010.510068
Markus, K. A. (2021). Causal effects and counterfactual conditionals: contrasting Rubin, Lewis and Pearl. *Economics & Philosophy*, *37*(3), 441–461. https://doi.org/10.1017/S0266267120000437
Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology*, *4:169*. https://doi.org/10.3389/fpsyg.2013.00169
McCullagh, P. (2002). What is a statistical model? *Annals of Statistics*, *30*(5), 1225–1267. https://doi.org/10.1214/aos/1035844977
McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
McDonald, R. P. (2003). Behavior domains in theory and in practice. *Alberta Journal of Educational Research*, *49*(3), 212–230. https://doi.org/10.11575/ajer.v49i3.54980
McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *the Annals of Probability*, *2*(4), 620–628. https://doi.org/10.1214/aop/1176996608
Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*(2), 105–118. https://doi.org/10.3102/10769986007002105
Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5
Mellenbergh, G. J. (1994a). Generalized linear item response theory. *Psychological Bulletin*, *115*(2), 300–307. https://doi.org/10.1037/0033-2909.115.2.300
Mellenbergh, G. J. (1994b). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*(3), 223–236. https://doi.org/10.1207/s15327906mbr2903_2
Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*(3), 293–299. https://doi.org/10.1037/1082-989x.1.3.293
Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/bf02294825
Michell, J. (2004). Item response models, pathological science and the shape of error: Reply to borsboom and mellenbergh. *Theory & Psychology*, *14*(1), 121–129. https://doi.org/10.1177/0959354304040201
Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling*, *2*(3), 255–273. https://doi.org/10.1080/10705519509540013
Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, *2*(3), 248–260. https://doi.org/10.1037/1082-989x.2.3.248
Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*(4), 461–473. https://doi.org/10.1007/s11336-007-9039-7

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. New York: Routledge. https://doi.org/10.4324/9781315871691

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. The Hague, the Netherlands: Mouton. https://doi.org/10.1515/9783110813203

Molenaar, I. W. (1995). Some background for item response theory and the rasch model. In F. G.H. & M. I.W. (Eds.), *Rasch models* (pp. 3–14). New York, NY: Springer. https://doi.org/10.1007/978-1-4612-4230-7_1

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*(3), 391–411. https://doi.org/10.1007/bf02296153

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551–560. https://doi.org/10.1007/bf02293813

Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in mplus. *Mplus web notes*, *4*, 1–22.

Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. New York, NY: Routledge. https://doi.org/10.4324/9781315871318

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*(1), 1–13. https://doi.org/10.1007/BF02289400

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146. https://doi.org/10.1214/09-ss057

Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 68–91, Chapter 5). New York, NY: Guilford Press. https://doi.org/10.21236/ada557445

Pohl, S., Steyer, R., & Kraus, K. (2008). Modelling method effects as individual causal effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(1), 41–63. https://doi.org/10.1111/j.1467-985x.2007.00517.x

Popper, K. R. (1959). The propensity interpretation of probability. *The British Journal for the Philosophy of Science*, *10*(37), 25–42. https://doi.org/10.1093/bjps/x.37.25

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Paedagogiske Institut.

Raykov, T. (2001). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*(1), 69–76. https://doi.org/10.1177/01466216010251005

Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9780203930687

Reichardt, C. S. (2006). The principle of parallelism in the design of studies to estimate treatment effects. *Psychological Methods*, *11*(1), 1–18. https://psycnet.apa.org/doi/10.1037/1082-989X.11.1.1

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rozeboom, W. W. (1966). *Foundations of the theory of prediction*. Homewood, IL: Dorsey Press.

Schmitt, M. (2006). Conceptual, theoretical, and historical foundations of multimethod assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 9–25). Washington, DC: American Psychological Association. https://doi.org/10.1037/11383-002

Schmitt, M., & Steyer, R. (1993). A latent state-trait model (not only) for social desirability. *Personality and Individual Differences*, *14*(4), 519–529. https://doi.org/10.1016/0191-8869(93)90144-r

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology*, *34*(2), 133–166. https://psycnet.apa.org/doi/10.1111/j.2044-8317.1981.tb00625.x

Shealy, R., & Stout, W. (1993). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.

197–240). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9780203357811

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. https://psycnet.apa.org/doi/10.1007/s11336-008-9101-0

Sijtsma, K., & Van der Ark, L. A. (2015). Conceptions of reliability revisited and practical recommendations. *Nursing Research*, *64*(2), 128–136. https://doi.org/10.1097/nnr.0000000000000077

Sijtsma, K., & Van der Ark, L. A. (2020). *Measurement models for psychological attributes*. Boca Raton, FL: Chapman & Hall/CRC. https://doi.org/10.1201/9780429112447

Silk, A. J. (1977). Test-retest correlations and the reliability of copy testing. *Journal of Marketing Research*, *14*(4), 476–486. https://doi.org/10.1177/002224377701400405

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*(2), 229–239. https://doi.org/10.1111/j.2044-8317.1974.tb00543.x

Spearman, C. (1904). "general intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state–trait theory and research in personality and individual differences. *European Journal of Personality*, *13*(5), 389–408. https://doi.org/10.1002/(sici)1099-0984(199909/10)13:5<389::aid-per361>3.3.co;2-1

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. https://doi.org/10.1007/bf02294363

Thielemann, D., Sengewald, M.-A., Kappler, G., & Steyer, R. (2017). A probit latent state irt model with latent item-effect variables. *European Journal of Psychological Assessment*, *33*(4), 271–284. https://doi.org/10.1027/1015-5759/a000417

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577. https://doi.org/10.1007/bf02295596

Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. Chicago, IL: University of Chicago Press. https://doi.org/10.1037/10018-000

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, *19*(5), 579–599. https://doi.org/10.1177/0959354309341926

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, *11*(1), 1–13. https://doi.org/10.1007/bf02288894

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*, 39–55. https://doi.org/10.1111/j.2044-8317.1990.tb00925.x

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer Science & Business Media. https://doi.org/10.1007/978-1-4757-2691-6

Van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339–356. https://doi.org/10.1037/a0022749

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength and Conditioning Research*, *19*(1), 231–240. https://doi.org/10.1519/15184.1

White, H. (2001). *Asymptotic theory for econometricians, revised edition*. San Diego, CA: Academic Press.