Chapter 5
Psychological Constructs as Organizing Principles

Denny Borsboom

Abstract Klaas Sijtsma has suggested that psychological constructs, such as those invoked in the study of intelligence, personality, and psychopathology, should be understood as *organizing principles* with respect to elements of behavior, including item response behavior. In a discussion in the journal *Psychometrika*, Sijtsma (*Psychometrika, 71*(3), 451–455 (2006)) contrasted this position with the *common cause* interpretation of Item Response Theory (IRT) models and the associated theory of validity that I had articulated some years earlier (Borsboom, *Psychological Review, 111*(4), 1061–1071 (2004)), arguing that this theory of validity was far too strong given the immature status of psychological constructs. In the present chapter, I present an alternative understanding of IRT models in terms of psychometric networks, which is inspired by Sijtsma's idea of constructs as organizing principles. From the weak premise that psychological constructs organize behaviors, in the sense of identifying behavioral elements that structurally hang together, in the present chapter, I show how one can build up a psychometric approach that can motivate and guide the use of tests in psychology in the absence of strong common cause interpretations.

5.1 Introduction

Psychometrics is an intrinsically multidisciplinary project, and like all multidisciplinary projects, it tends to disintegrate into unconnected monodisciplinary components if left to its own devices. Klaas Sijtsma is one among a small group of psychometricians who have spent their careers trying to protect the brittle but essential connections between substance, mathematics, and philosophy. In this respect, Klaas and I are kindred spirits, because both of us have tried to find a balance between the messy reality of psychometric practice, the idealized structures

D. Borsboom (✉)
Department of Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands
e-mail: D.Borsboom@uva.nl

L. A. van der Ark et al. (eds.), *Essays on Contemporary Psychometrics*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-031-10370-4_5

of psychometric modeling, and the conceptual questions of what psychological measurement is and how it should be optimized.

Despite these shared commitments, in the past, Sijtsma and I also have defended different positions on the core question of how psychometrics should relate to its neighbors. Sijtsma has argued that psychometrics should operate as an auxiliary discipline to psychology, i.e., it should seek a partnership in which it plays the role of helper (Sijtsma 2006). I have tended to take a more directive position, primarily because I doubt that psychologists are sufficiently interested in measurement to tackle the problems involved (Borsboom 2006; Borsboom et al. 2004).

Unfortunately, the theoretical basis required for the research program I championed (Borsboom et al. 2004) is often unattainable in psychology, as Sijtsma (2006) astutely observed, because standard measurement models in psychometrics are unrealistic given the substance matter of psychology. In recent work, however, alternatives to standard measurement models have been developed that seem to align much more naturally to the way that psychologists think; in these models, constructs are not seen as common causes of manifest variable, but as network structures that connect such variables (Borsboom et al. 2021). It turns out that these models are actually finely tuned to a comment that Sijtsma (2006) made in discussion we had in *Psychometrika*, in which he presented the viewpoint that psychological constructs should operate as "organizing principles" that specify which psychometric items "hang together."

In this chapter, I aim to bring this idea of Sijtsma (2006) in contact with the field of network psychometrics, which has been recently developed on the basis of the network perspective on psychometric constructs (Marsman et al. 2018; Borsboom et al. 2021; Van Borkulo et al. 2014; Cramer et al. 2010) to arrive at an alternative conceptualization of psychometrics in the context of network models. I first review the standard interpretation of latent variables as common causes, after which I discuss an alternative interpretation in terms of structurally connected variables. Finally, I examine the important psychological concepts of unidimensionality, reliability, and validity from this viewpoint.

5.2 Item Response Theory and Common Cause Structures

Item Response Theory (IRT) models the response of a person i to an item j as a function of a set of item and person parameters through an Item Response Function (IRF) that maps each combination of the parameters to a probability distribution over the item responses. In the case that there is only one person parameter θ_i, we have a unidimensional model. A commonly used example of such a model is the well-known Rasch (1960) model, in which the IRF is logistic and each item has one parameter, β_j, which controls the location of the IRF:

$$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}} \tag{5.1}$$

Because of its ease of application, mathematical tractability, and favorable measurement properties, the Rasch model is popular among psychometricians. It is heavily used in fields like educational testing, intelligence and personality research, and the study of psychopathology. The model will therefore serve well as a leading example in the current chapter.

Looking at the Rasch model, it is evident that the item response probabilities are the result of a trade-off function between the item and person parameters, which are often called “difficulty” and “ability,” reflecting the origin of the model in educational measurement. This trade-off is possible because θ and β are on the same scale, which means that “difficulty” and “ability” are, in an important sense, exchangeable: “having a higher level of ability” is equivalent to “making an easier set of items,” not just in a figurative mode of speech, but exactly. The fact that all of the IRFs that describe a set of items are controlled by a single person parameter then means that each of the item difficulties trades off against the same ability. This, in turn, suggests that θ functions as a *common cause* of the item responses (Reichenbach 1956; Pearl 2009; Haig 2005a,b).

It is useful to briefly consider the notion of a common cause, as introduced by Reichenbach (1956), to establish this parallel. Reichenbach (1956) dealt with the situation in which a binary common cause, C, has two binary events A and B as its effects. In this case, a common cause is required to satisfy three conditions: (1) $P(A|C) > P(A|\neg C)$ and $P(B|C) > P(B|\neg C)$, (2) $P(A \cap B) > P(A)P(B)$, and (3) $P(A \cap B|C) = P(A|C)P(B|C)$. A classic example considers the relation between yellow-stained fingers (A) and lung cancer (B) as a function of smoking (C): the probability of both yellow-stained fingers and lung cancer is increased, given smoking (condition 1) yellow-stained fingers and lung cancer are positively associated (condition 2), and smoking “screens off” the association between yellow-stained fingers and lung cancer, rendering them conditionally independent (condition 3).

Translating this to a situation with m dichotomous effect variables X_j, $j = 1, \ldots, m$ and a continuous common cause θ, as would match most IRT models, Reichenbach’s conditions become:

1. $P(X_j = 1|\theta)$ is increasing in θ.
2. $P(X_j = 1, X_k = 1) > P(X_j = 1)P(X_k = 1)$ for all j, k.
3. $P(x_1, \ldots, x_j, \ldots, x_m|\theta) = \prod_{j=1}^{m} P(x_j|\theta) = \prod_{j=1}^{m} P(X_j = 1|\theta)^{x_j} P(X_j = 0|\theta)^{1-x_j}$.

Condition 1 is satisfied in the Rasch model, as the logistic function (1) is strictly increasing in θ. Condition 2, positive association, is a well-known consequence of every unidimensional monotone latent variable model (Holland & Rosenbaum

1986) including that of Rasch. Condition 3 is local independence, a common property of IRT models, including that of Rasch. Thus, the Rasch model conforms to a common cause structure.

In fact, conditions 1–3 are satisfied in all unidimensional models for dichotomous item responses that have increasing IRFs, like the popular model of Birnbaum (1968). In less restrictive models, like the Mokken (1971) nonparametric model and its generalization, the monotone latent variable model (Holland & Rosenbaum 1986; Junker & Sijtsma 2001), a weaker form of monotonicity (i.e., that $P(X_j = 1|\theta)$ is non-decreasing in θ) exists that does not strictly conform to these conditions; however, in such models, the latent variable can be conceived of as the common cause of subsets of item responses, in those regions of θ where the corresponding items' IRFs are all increasing. Thus, Reichenbach's (1956) common cause structure applies to the relation between θ and the item responses in a broad class of IRT models.

This appears to be more than a statistical coincidence, because several other psychometric concepts have strong parallels with the causal modeling literature as well. For instance, in a measurement context, it is sensible to require that θ mediates the effects of a set of external factors $\{V\}$ on the set of items $\{X\}$. That is, if $\{X\}$ measures θ, then changes in the item response probabilities induced by conditioning on group variables (e.g., sex) or interventions (e.g., therapy) should affect the item responses only indirectly, that is, through θ. In causal terms, this means that θ should "block" all causal paths from variables in $\{V\}$ to variables in $\{X\}$. Via the criterion of *d-separation* (Pearl 2009), this implies the following conditional independence relation for all variables in $\{X, V\}$:

$$F(x|\theta) = F(x|\theta, v), \tag{5.2}$$

for all (θ, v), where $F(x|\theta, v)$ denotes the value of the conditional distribution function of X evaluated at the point (θ, v). In the psychometric literature, (2) is well known as the requirement of *measurement invariance* (Mellenbergh 1989; Meredith 1993; Millsap 2007). Interpreted causally, measurement invariance thus requires that no variables except for θ exert a direct causal effect on the item responses.

The idea that θ acts as a common cause of the item responses also matches the way many substantive researchers think about latent variables. Spearman (1904) set up the common factor model to analyze cognitive tests in accordance with this notion, as he interpreted general intelligence, or g, as a source of individual differences present in a wide range of cognitive tests (see also Jensen (1999), for a similar view); the condition of *vanishing tetrads* that Spearman introduced as a model test is currently seen as one of the hallmark conditions of the common cause model (Bollen & Ting 1993). In personality research, putative latent variables such as those in the Five-Factor Model are likewise seen as causes of behaviors; for instance, McCrae and Costa Jr. (2008) argue such things as "E[xtraversion] causes party-going" (p. 288). Finally, in clinical psychology, Reise and Waller (2009) note that "to model item responses to a clinical instrument [with IRT], a researcher must first assume that the item covariation is caused by a continuous latent variable" (p. 26).

Thus, not just the "letter" (i.e., the formal correspondence given above) but also the "spirit" of latent variable modeling is driven by the idea that our item responses are the effects of a common attribute that underlies the observations, represented in the model structure by the symbol θ. As Reise and Waller (2009) note, this "sets limits on the type of constructs that can be appropriately modeled by IRT" (p. 26); namely, the type of constructs for which this is sensible is the type for which, minimally, it can be expected that the items will behave as if they are a function of a common cause.

5.3 The Causal Account of Test Validity

The common cause understanding of latent variable models is strong but clear. In 2004, I developed a straightforward consequence of the causal interpretation of measurement models for the concept of validity (Borsboom et al. 2004). My reasoning was that, *if* psychological constructs like depression or intelligence signify common cause of test scores, *and* validity refers to the question of whether these test scores measure what they should measure, *then* the core of any validity argument must lie in specifying the psychological processes by which the relevant psychological attributes play their causal role. This idea applies naturally for certain test types; an example may involve items as used in working memory capacity tests. In these tests, participants are instructed to recall different sequences of letters or numbers, while they are simultaneously executing another task (e.g., counting back from 100 to 0). Plausibly, one's success in recalling the sequence 2, 6, 4, 7, 2 and the sequence 4, 6, 3, 8, 9, 4, 3, 4, 5 depend on the same resource, namely, working memory capacity. Clearly, then, working memory capacity acts as a common cause with respect to the individual differences in item responses.

This type of causal argument says how individual differences in a psychological attribute, which affects all of the item responses, are translated into individual differences in test scores. In my view, this forms the core of the validity concept. If one thinks about it, such specifications are not hard to come by in cases where questions of validity actually have a definite answer. Such examples, in my view, are too scarcely considered in validity theory. In fact, the idea that validity questions are unanswerable is taken for granted in certain lines of thinking about validity (one received view is that "validity is a never-ending process"). However, there are actually measurement problems that have been solved and validity questions that have been answered. And typically, the answer to a question like "why does instrument X measure attribute Y?" hinges on a specification of *how the instrument works* (i.e., specifies a causal process where the measured attribute is the starting point and the meter readings are the endpoint). Why do mercury thermometers measure temperature? Because higher temperatures cause the mercury to expand and hence the meter rises. Why does the composition of air trapped in the Arctic ice measure historical global carbon dioxide emissions? Because higher emissions cause more carbon dioxide in the relevant air pockets and higher concentrations of carbon dioxide cause higher readings in spectral analysis of the air contained in these pockets. Why does the item "what is your age?" measure age? Because people know how old they are and, if willing, will be able to supply that information.

If available, the causal answer to validity questions is the most forceful answer there is. It is explicit, testable, and suggestive of changes that might improve the measurement device. However, it is also a very taxing answer. It requires a convincing account of how the measured attributes exert their causal effects, and theories that can motivate such accounts are scarce in psychology (although they do exist, as some of the above examples show).

In 2004, I believed that this type of analyses could be made to work in psychology at large and should be investigated vigorously. Our task as psychometricians, in my view, was to come up with good analyses of response behavior in which the measured attribute played a causal role. It looked like that kind of analysis was there for the taking with the combination of advanced modeling techniques, cognitive diagnostic models, and good psychological theory. However, some colleagues were skeptical. Klaas Sijtsma was one of them (Sijtsma 2006). In response to a paper in which I pushed the causal psychometric account to its extreme (Borsboom 2006), he articulated doubts with respect to the research program I was advocating:

> Borsboom's assumption about the ontology and causality of psychological attributes seems to lead to a very restrictive conception of the process of construct validation: Elegant in its rigor but impractical for psychology (and many others areas). It seems to me that we still know so little about the functioning of the human brain in general and cognitive processes including those underlying personality traits and attitudes in particular, that it is difficult even to say what an 'attribute' is. In the absence of such knowledge, I prefer to consider psychological attributes as organizational principles with respect to behavior. Thus, my point of view is that psychological attributes define which behaviors hang together well and are useful to the degree in which tests sampling these behaviors play a role in predicting interesting psychological phenomena.

With some reluctance, I have to admit defeat to this charge when it comes to the more abstract entities in the psychometric pantheon—that is, the big psychometric players like general intelligence, neuroticism, attitudes, and psychopathological conditions. In the years that followed the conceptual articulation of the causal validity program, I attempted to come up with good measurement theories for such constructs but ultimately failed to provide a believable analysis in causal terms. Although this research line of mine is undocumented and impossible to replicate—a failure to construct conceptual analyses leads to the theoretical equivalent of a file-drawer problem; one can hardly publish failures to come up with a new theory—I did try hard. Apart from a few isolated successes (most notably the analysis of IRT model results in terms of drift diffusion parameters as developed by my colleague Han van der Maas (Van der Maas et al. 2011)), it just didn't work.[1]

[1] Naturally, that I could not come up with good theories of test validity does not mean that nobody else could. Perhaps I didn't use the right framework; perhaps I just approached the problem from

In fact, that is an understatement. If one attempts to specify how general intelligence causes responses to the item “Who wrote the Iliad?”, how depression leads to sad mood, and how attitudes influence the answer to questions like “do you think Trump is a good leader?”, one arrives at theories that are far too strong and far too simplistic. In fact, the very idea that traits like intelligence, extraversion, and psychopathological syndromes are causes of human behavior, including the behavior that involves ticking boxes on questionnaires, appears to be rather far-fetched, more akin to Moliere’s *virtus dormitiva* than to any serious appreciation of the psychological complexity of the constructs in question.[2] B.F. Skinner (1987) once stated that “as soon as you have formed the noun *ability* from the adjective *able*, you are in trouble,” and indeed that seems to be accurate for many of the abilities and traits invoked in psychometric theory.

5.4 Structural Connections

The general failure to come up with adequate measurement theories forms an interesting contrast with the relative ease with which one can concoct psychometric models. Taking desirable measurement properties as axiomatic for measurement models, it is possible to deduce the general form and structure that psychometric models *should* have and work out the distributions of data they imply. This is what, in my view, psychometricians have been most successful at over the course of the past century. One can easily imagine the tests and theories employed in psychology today to become a laughing stock for future generations, but the intricate building of interrelated statistical measurement models of IRT, which Klaas Sijtsma and others erected in the past decades, will remain an important entry in the scientific record.

Because such models have more to do with philosophical ideas on what good measurements should look like, than with psychological ideas about whatever it is we are measuring, psychometric models are in my view best seen as applied philosophy of science. The models one can deduce from general philosophical measurement desiderata range from very strong to extremely weak. The Rasch model in Eq. 5.1 is an example of a strong model. Rasch (1960) started from some desirable measurement axioms (e.g., things that would be nice to have, like separate identifiability of person and item characteristics) and then deduced the

the wrong angle; may others come and do it better. However, as they say, insanity is trying the same thing over and over again and expecting different results, so it seemed more sensible to reconceptualize my problems than to keep trying.

[2] As an aside, if test score use and interpretation would actually require theories of this kind, then the whole scientific project of psychometrics would be in serious trouble, perhaps even trouble of the end-of-story kind. Realizing this, in hindsight, it is unsurprising that the reception of my validity theory was mixed. One influential validity theorist stated informally that what I said might all be good and true, but that my definition of validity would never be accepted because theories that specify how psychological constructs cause item scores “would not hold up in court.”

model formula in Eq. 5.1 as a consequence. One can also proceed from much weaker requirements and deduce weaker models as a result (Holland & Rosenbaum 1986; Ellis & Junker 1997); this more realistic approach is the cornerstone of nonparametric IRT, to which Junker and Sijtsma (2001) and Sijtsma and Molenaar (2002) provide excellent introductions.

The focus on desirable measurement properties leads to simple models. The Rasch model in Eq. 5.1 is one example, but basically all models in the IRT family (Mellenbergh 1994) are variants of the general structure. Usually, that structure specifies how people's position on a relatively simple latent variable (e.g., a point on a continuous line, membership of a latent category) is coordinated with a specific probability distribution over the item responses. Because nearly all models specify a form of conditional independence, in which the observed variables are independent given the latent variable, they can typically be understood along the lines of Reichenbach (1956) as explained in the previous paragraph. Thus, nearly all models can be understood as specifying a (possibly somewhat convoluted) common cause model.

However, if we think for a moment about, say, relations between symptoms of depression, attitude items, or cognitive processes, it is hard to see how causal interpretations of such simple models could possibly be on target. After all, it would be a small miracle if human behavior, embedded in a nexus of complex interactions between factors at genetic, physiological, psychological, and social levels, were literally governed by a model structure as simple as Eq. (5.1) and its relatives.

This realization, however, presents us with a paradox. This is because the latent variable modeling approach in topics, like intelligence, personality, and psychopathology, has *not* fared as badly as one should expect, given the complexity of human behavior. Although measurement models rarely fit adequately, they do generally provide a reasonable description of the data; for instance, the fact that the general factor of intelligence is now in the company of general factors of personality and psychopathology is not accidental. In recent years, I have investigated the hypothesis that the reason for this is that the tests used in such domains depend on distinct attributes and processes that do *not* depend on a common cause, but *are* structurally connected through relations that can reasonably be approximated by pairwise interactions; these pairwise interactions, in turn, generate probability distributions that tend to fit latent variable models reasonably well.

What does it mean for variables to be structurally connected? To preempt some obvious misinterpretations, let me first say what I do not mean. First, I do *not* mean to say that structurally connected variables merely correlate. Ice cream consumption and murder rates are famously correlated across the months of the year, but not structurally connected. Second, to be structurally connected does not necessarily mean that variables stand in directed causal relations. Sad mood and suicidal ideation, for instance, are probably to some extent involved in some reciprocal reinforcement process, but it is unlikely that this relation is of the smoking-causes-lung-cancer kind that modern theories of causality (e.g., Pearl (2009)) present as axiomatic. In addition, I intentionally cover cases where different items are in part related through semantic or logical pathways. For example, some items in personality questionnaires contain very similar wordings, which leads responses to be structurally connected, but the queried attributes are unlikely to stand in directed causal relations.

As a working definition, I propose variables to be structurally connected if they (or their probability distributions) cannot vary independently. This definition is extremely broad and covers a wide variety of cases where relations between variables are systematic (i.e., they are not merely correlated) but not necessarily causally directed. For variables to be structurally connected thus means that these variables represent elements of behavior that, in the words of Sijtsma (2006), "hang together."

Here are some examples. Responses to the item "do you think Trump is a good leader?" are structurally connected with responses to the item "do you like Trump?" because people strive to keep their attitude elements consistent. Responses to the item "do you like parties?" are structurally connected with responses to the item "did you like the last party you went to?" because the latter assesses a memory trace that a respondent will also use in answering the former. Responses to the item "have you felt fatigued over the past 2 weeks?" are structurally connected with responses to the item "have you slept more than usual over the past 2 weeks?" because people who are tired will tend to sleep more. Responses to the item "have you felt fatigued over the past 2 weeks?" are *also* structurally connected with responses to the item "have you slept less than usual over the past 2 weeks?" because people who don't sleep well tend to get tired. In each of these cases, the relevant variables cannot vary independently, because they share meaning, are causally related, share resources, or are intertwined in development.[3] In contrast, responses to the item "do you like parties?" are not structurally connected with responses to the item "who wrote the Iliad?", because these variables can vary independently. For the same reason, responses to the item "have you felt fatigued over the past 2 weeks?" are not structurally connected with responses to the item "do you think Trump is a good leader?".

5.5 Network Representations of Psychological Constructs

Shifting attention from a common cause principle to the idea of structural connections between variables invites a different way of setting up our basic psychometric apparatus. I propose to denote the structural connection between two variables X_j and X_k with a tilde:

$$\text{Structural connection} \equiv X_j \sim X_k \tag{5.3}$$

[3] In a very weak interpretation of causality, one could say structural connections are a type of causal relations, but I think this stretches the meaning of the term beyond the limits of usefulness.

$X_j \sim X_k$ means that the variables in question cannot vary independently. One way of making this idea more precise might be taken by saying that intervening on X_j will affect X_k and vice versa, i.e., implying a bidirectional causal relation between the variables in question that can be expressed using the concept of a Do-operator (Pearl 2009). The Do-operator is used in the causality literature to represent interventions on a system in order to provide a semantics for causal relations. In particular, a causal effect of X_j on X_k would be expressed as $P(X_k|\mathrm{Do}(X_j = x_j)) \neq P(X_k)$, i.e., a causal effect means that the probability distribution of X_k is not the same under manipulations that force X_j to take different values x_j. In the present case, one could imagine that a structural connection may be taken to imply bidirectional causal dependence:

$$X_j \sim X_k \Rightarrow P(X_j|\mathrm{Do}(X_k = x_k)) \neq P(X_j) \wedge P(X_k|\mathrm{Do}(X_j = x_j)) \neq P(X_k) \tag{5.4}$$

This type of characterization in causal terms may be useful to flesh out specific formalizations of structural dependence.[4] For instance, given the causality calculus, the causal formulation implies the statistical consequence that two variables cannot be rendered statistically independent, given any other variable at our disposal. Thus, given a set of variables $\{X\}$ that characterize a system under study, if a structural connection exists between X_j and X_k, this implies that when conditioning on the complement set $\{X_c\}$ (all variables in $\{X\}$ excluding X_j and X_k):

$$X_j \not\perp\!\!\!\perp X_k|X_c \tag{5.5}$$

In other words, the variables are not statistically independent given everything else we can measure on the system. Ordinarily, the set $\{X\}$ will be a pragmatically chosen collection of variables, and the question of whether any two variables are structurally connected is studied relative to this collection. It would be interesting to investigate what other choices would be sensible to define the set $\{X\}$ or what it means for $\{X\}$ to characterize the system under study, but I will not pursue these questions here and will simply assume $\{X\}$ to be composed of whatever a researcher chooses to include in the data. Also, for convenience, I will assume the bidirectional causal relation to be symmetric (i.e., equally strong in both directions) although I don't think much hinges on that.

For a given set of variables, the above definitions imply a network structure to which standard representations of network psychometrics apply. In particular,

[4] I hasten to add that the above characterization should be seen as one of the various ways to make the idea of a structural connection concrete and not as definitional. Also I do not intend the notion of structural connection to require such things as decomposability (de Boer et al. 2021) and similar kinds of atomistic conception of the world of variables, which seems to have become common in the language of causality. Thus, the implication is not biconditional as the causal analysis does not exhaust the possibilities and may depend on auxiliary assumptions that are not satisfied in psychometrics.

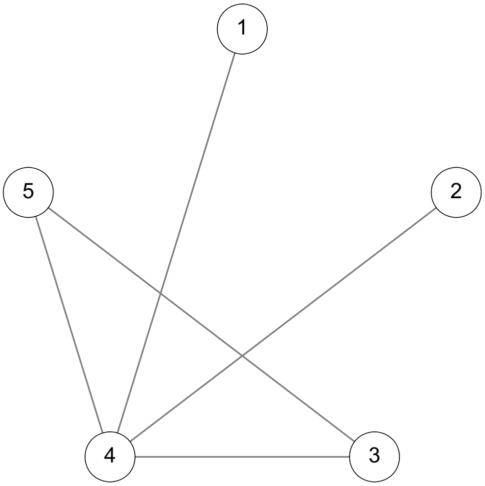

Fig. 5.1 A network of five variables. Edges between variables indicate that the relevant variables are structurally connected

Eq. 5.5 defines a Pairwise Markov Random Field (PMRF), which has the attractive graphical representation as a network in which variables that are not directly connected are conditionally independent given the other variables. For binary variables, the PMRF can be estimated in various ways, for instance, through the R-package IsingFit (Van Borkulo et al. 2014).

Figure 5.1 provides an example network. Variables are represented as nodes, and structural connections as edges. The set of nodes that is connected to node j is known as the *neighborhood* of j and denoted N_j. We assume that the probability distribution of the variables has the Markov property, i.e., that it factorizes according to the graph structure. This implies that the joint probability distribution can be represented as log-linear model that includes main effects for all variables and pairwise interactions for any two variables that are connected in the graph. However, for my current purposes, it is more convenient to think of the model in terms of a set of logistic regressions, where each node is regressed only on the variables in its neighborhood:

$$\text{logit}(X_j) = \alpha_j + \sum_{k \in N_j} \beta_{jk} X_k \tag{5.6}$$

This formulation is the model used in the IsingFit representation (Van Borkulo et al. 2014). Now let us consider the relation between Eq. 5.6 and the typical IRT representation as in Eq. 5.1. Willem Heiser (personal communication) has observed that this representation connects the network approach to an older tradition in psychometrics, namely, that of image factor analysis (Guttman 1953). Image factor analysis is an approach to factor analysis that explicitly aims to avoid the use of latent variables. In image factor analysis, the regression of a variable on all other variables in the data creates the variable's image (the weighted sumscore formed by the regression), while the residual of that regression defines its anti-image. The

IsingFit representation of the network model could be seen as an extension of the image factor analysis model to the dichotomous case; in this interpretation, the regression model defines the image of X_j (Guttman 1953).

5.6 Reinterpreting Psychometric Concepts

In the IRT model, the items are related to a single person parameter, and the regression parameter for that function depends only on the item considered. In the logistic regression formulation of the network model, the items are related to a set of independent variables, and the regression parameters are different for each of them. In the IRT formulation, the predictor is latent. In the logistic regression formulation, it is observed.

However, there are also similarities. In both cases, we see a generalized regression with a parameter that depends on the item (the intercept in the logistic regression, the difficulty parameter in the IRT model) and a regression parameter that controls the slope of the regression of the item on the predictor term. That predictor, in the IRT model, is the latent variable. In the logistic regression, it is a set of scores on the neighboring items. These scores are weighted by regression weights. We can imagine collecting the combined effects of all predictors in a weighted sumscore of the variables in the item j's neighborhood, which for person i we may denote as

$$N_{ij}^{+} = \sum_{k \in N_j} \beta_{jk} X_{ik} \tag{5.7}$$

Now things start to look quite analogous if we express person i's expected score as a function of the latent variable model,

$$P(X_{ij} = 1 | \theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}, \tag{5.8}$$

and as a function of the regression model. We can make this similarity most apparent by putting the regression in the same form as the IRT model through suitable transformations of parameters, representing the model in terms of a trade-off between the internal field (the effects of the other nodes in the network) and the external field (e.g., in case the regression coefficients equal unity, this would directly correspond to the intercept parameter in Eq. 5.6 transformed to $\alpha_j^* = -\alpha_j$):

$$P(X_{ij} = 1 | N_{ij}^{+}, \alpha_j^*) = \frac{e^{N_{ij}^{+} - \alpha_j^*}}{1 + e^{N_{ij}^{+} - \alpha_j^*}} \tag{5.9}$$

Using this representation, we see that the neighborhood score N_{ij}^{+} plays a role that is analogous to that of the latent variable θ in the IRT model, while the intercept of the regression α_j^* is the analogue of item difficulty in IRT. Via the concept of structural connection, one can think of any specific item as standing under the influence of the variables in its neighborhood, not in the sense that its value is directly caused by these, as in a billiard ball causation picture, but in the sense that the item's probability distribution cannot change independent of that of its neighbors. In a nontrivial sense, therefore, the item *measures* the influence of its neighbors: ceteris paribus, the more neighbors of item j are positive (take value $X = 1$), the more j will tend to be positive as well.[5]

The relation between the latent variable in IRT and the neighborhood score in network analysis in the dichotomous case mirrors the relation between latent variables in factor analysis and components in image factor analysis for the continuous case (Guttman 1953); also, the centrality measure of predictability that has been proposed in the network literature (Haslbeck & Waldorp 2018) is highly similar to the index of determination discussed in Guttman (1953). Finally, note that the dimensionality of the neighborhood scores is the same as that of the data (i.e., there are as many neighborhood scores as variables); a reduction of these neighborhood scores could be achieved through, for instance, a principal component analysis, which would compress the neighborhood scores into a smaller dimensionality. In the case where the network is fully connected, one would then expect the neighborhood scores to approximate unidimensionality, while a sparsely connected network would not.

Although the alignment between IRT and network models that I have constructed here is not as mathematically elegant as those used in the direct equivalence proofs between multidimensional IRT and Ising models that are now in the literature (Marsman et al. 2018; Epskamp et al. 2018), the logistic regression of an item on a neighborhood score has intuitive appeal and facilitates reinterpretation of psychometric concepts. This is because we can keep in mind the analogy between the latent variable and the neighborhood score. Substituting the concept of a neighborhood score in a network of structural connections for the concept of a common cause of item responses leads to several straightforward consequences for psychometric practice. In the following, I review some of the most important psychometric concepts from this point of view.

[5] The relation between the item and the targeted latent variable is typically represented in an Item Characteristic Curve (ICC). Of course, we can do the same in the network model, if we put N_{ij}^{+} on the x-axis and the probability of a positive item response on the y-axis; we may call this curve a Network Response Function (NRF). The items will have different neighborhoods, which means the NRFs have different domains, but the general concept clearly is similar.

5.6.1 Unidimensionality

The notion of unidimensionality plays a very important role in psychometrics. It encodes the idea that the correlations between item responses can be represented as a function of a single dimension. In the parlance of IRT, unidimensionality means that the logits of the expected scores of the items (the true item scores) are perfectly correlated, which means that if one knows a person's true score on one item, one cannot learn anything new about the ordering of the persons on the latent variable by consulting the other items. In terms of the causal interpretation of measurement models, this represents the hypothesis that the different items trade off against precisely the same ability.

Networks in general are unlikely to satisfy such requirements, but they can approximate them (and often do). This works as follows. If one looks at Fig. 5.1, it is clear that the variables have very different neighborhoods. Node 1 only has one neighbor (node 4), while node 4 has four (nodes 1, 2, 3, and 5). Clearly, in this case, the covariance matrix will depart from unidimensionality significantly. However, if one imagines an ever more densely connected network, one can see that the neighborhoods of different nodes will overlap more and more. Thus, the neighborhood scores of different items will get more and more correlated. In a perfectly connected network, the neighborhoods of any two nodes will differ by only one term (the scores on the evaluated nodes themselves, which are not part of their own neighborhood). Thus, the closer the network approaches perfect connectivity, the closer it will get to unidimensionality. In the network literature, this means that the network can be approximated by the so-called mean field approximation, which essentially substitutes a single number for all of the node neighborhoods (Finnemann et al. 2021). In a nontrivial sense, the latent variable in a unidimensional psychometric model corresponds to the mean field in a network model, which in turn is strongly related to the first factor of an image factor analysis (Guttman 1953).

One can also see that, as the network gets larger, the neighborhoods get ever more close. I conjecture that this, in effect, realizes the same process that Ellis and Junker (1997) describe through the concept of a tail measure. A tail measure is the equivalent of a sumscore on an infinite item domain, which Ellis and Junker (1997) showed is an adequate interpretation of a latent trait. Similarly, I suggest that an infinitely large network will produce equivalent tail measures on items' neighborhood scores, as in the limit all neighborhoods will coincide in terms of their ordering of persons. Thus, from a network perspective, unidimensionality can be interpreted as a measure of network homogeneity. Interestingly, a perfectly connected network with invariant edge weights (a so-called Curie-Weiss model) turns out to be statistically equivalent to the Rasch model (Marsman et al. 2018).

5.6.2 Reliability

It would not be much of an overstatement to say that Classical Test Theory (CTT) was invented to furnish a basis for the notion of reliability: the degree to which true scores are linearly predictable from observed scores. The most important estimator of reliability, Cronbach's α, is controversial in psychometrics, both because of misinterpretations of the concept and because it is statistically inferior to other estimators (Sijtsma 2009). However, it is probably also the most important quantity psychometrics has delivered, as it regulates the composition and size of item sets used in practical test applications.

Reliability is commonly seen as a property of a test.[6] That is, it is a measurement concept, which indicates to what extent the total test score contains "measurement error." However, it is a well-kept secret among psychometricians that the noise in our test scores is rarely identifiable as measurement error independent of the psychometric model. Typically, what we call measurement error is simply variance that simply cannot be explained from the latent variable model (for whatever reason). Why this unexplained variance should be interpreted as measurement error is rarely explicated.

Interestingly, in the network representation, the psychometric representation of sumscore reliability is not (only) a measurement concept. Even if all items are measured without error, the network may still leave variance unexplained, for instance, because the items do not hang together perfectly (i.e., there is wiggle room for individual items given the other items) or because the network is not fully connected. This may very well be a property of a *construct* rather than of the *measurement instrument*. Indeed, Dalege and van der Mass (2020) hold that implicit measures of attitudes are necessarily unreliable because in the situation where people do not attend to the attitude, the attitude network operates in a high entropy regime (i.e., the network is weakly connected).

What *does* reliability imply, from a network perspective? In my view, high reliability means that the state of the individual items is highly predictable from the neighborhood scores. That is, the network has a low entropy (Dalege & van der Maas 2020), because the structural connections between items are strong so that items tend to align. Interestingly, low entropy implies that more extreme sumscores will become more prevalent, which will lead to higher variance of the sumscore. Thus, from a network perspective, the ratio of the sum of the item variances to the total test score variance—a standard operationalization of reliability—is actually a measure of how strongly connected the network is.

[6] This is fundamentally mistaken because, even on its own terms, CTT represents reliability as a test×population interaction (Mellenbergh 1996), but I will ignore this here and assume the population given.

5.6.3 Validity

As I noted earlier, in the past, I have articulated and defended the idea that validity is a causal concept, which hinges on the degree to which the measured attribute (represented as a latent variable) influences the item scores. Clearly, in the network representation, there is no latent variable (except as a mathematical representation of the joint probability distribution of the network; Epskamp et al. (2018)). Hence, the causal interpretation of validity is not on offer for the network as a whole. However, that conception can still be operational for the individual items in a network, for instance, if one asks whether the depression item "have you slept less than usual over the past 2 weeks?" actually measures insomnia (Cramer et al. 2010). In addition, if different items depend on a variable that is not represented in the network (i.e., a latent variable), then a latent variable model can be used to analyze that part of the network (e.g., in a latent network model; see Epskamp et al. (2017)), and in this case, the latent variable can be conceptualized as a common cause, which renders the causal account of validity applicable.

But what can one say about the validity of a test if the items in that test in fact measure properties that are structurally connected, rather than a single latent attribute? If the network model is true, then the construct label (e.g., "depression," "intelligence," "neuroticism") does not refer to such a latent attribute but to the network as a whole. Thus, when we ask "does this depression questionnaire actually measure depression?", the question should be understood as "do the variables assessed through the items included in this questionnaire actually correspond to the nodes in the depression network?". This, in turn, leads to the question "which nodes are part of the depression network?". And it is here, I submit, that the psychometric construct fulfills its function as an *organizing principle*. A construct label such as "depression" does not designate a latent attribute targeted in the measurement procedure, but instead indicates a family of variables that are structurally connected to produce the coordinated behavior of the network as a whole that we phenomenologically recognize as the overall state of individuals we are interested in.

Thus, the organizing principle of psychological constructs involves a simple but important task: to identify which nodes should be part of the network. In the special case that items are questions (rather than observations of behavior or other modes of investigating the human system, such as brain states or genetic profiles), this means that psychological constructs fulfill their main function in the area traditionally referred to as *content validity*. This is ironic, because in the literature on validity theory, content validity is typically seen as an outdated concept, if not an inferior one (Guion 1980). If the combination of Sijtsma's "hanging together" and network psychometrics is in the right ball park, content validity may thus well see a revival.

5.7 Discussion

In the present chapter, I have offered a reinterpretation of standard psychometric concepts in terms of a network perspective, in which item responses are viewed as structurally connected components in a network. This perspective aligns remarkably well with the idea that item responses merely "hang together" (Sijtsma 2006). In the presented scheme, the role of the psychological construct is radically shifted: the construct label does not designate a latent variable that acts as a common cause with respect to the item responses, but a set of relevant properties that are structurally connected. The primary task of the construct theory, so understood, is to indicate *which* of the many potentially relevant properties actually is part of the psychometric network, i.e., is part of the set of structurally connected variables.

This is quite a different way of thinking about the function of construct theories, but it seems to fit psychological practice quite well. Whenever I proposed to substantive psychologists that their theories should provide information with respect to the question of how a latent attribute determined the responses to questionnaire items, they looked at me as if they witnessed water burning. However, most of these same psychologists will have little problems in identifying why certain items should be included in a test. Usually, their answers either implicitly or explicitly explain how the items tap attributes that hang together systematically. The reasons behind these connections can vary wildly from area to area, so they cannot be uniformly fleshed out. However, in many cases, the connections in question suggest that variables bear a connection that is stronger than mere association and weaker than directional causation. I have tried to capture this notion in the term "structural connection."

My exploration of the mathematical conceptualization and the theoretical consequences of this idea has been preliminary. Especially the connection to the work by Ellis and Junker (1997) seems to harbor some interesting secrets that I have not developed here. In particular, because in a positive manifold that is consistent with a unidimensional factor model, pairwise conditional associations are always weaker than unconditional ones (Van Bork et al. 2018), it seems that in such cases the size of networks is limited by the strength of the structural connections that they consist of. That is, mathematically, a set of items that realizes an item domain can grow without bound (in fact, this is required for the proofs in Ellis and Junker (1997)). But a fully connected network like the Curie-Weiss model discussed in Marsman et al. (2018) cannot grow without bound, unless the conditional associations in the network get ever smaller in the process. It seems to me that this will not always be attainable. In other words, sets of items that cover a fully connected network may have a limited size. This would induce the notion of a construct that features a finite item domain which, to my knowledge, has not yet been developed in psychometrics.

As noted in the introduction to this chapter, psychometricians have many alliances. Their models, while cast in the language of mathematics, have an important connection to substantive realms (e.g., psychology, education, etc.) as well as to conceptual ideas about the nature of measurement. These alliances often clash. What is desirable from the point of view of measurement theory (e.g., additivity of the model, separability of parameters, simplicity, and parsimony) is often substantively speaking unrealistic. On the other hand, processes that are relevant from a substantive point of view (e.g., in terms of cognitive processes involved in psychometric tests) often lead to theoretical models that are mathematically intractable and that do not respect the strictures that the occupants of measurement theory's ivory tower proscribe as normative. The challenge is therefore to find conceptions of psychometric constructs that have a natural representation as mathematical structures, so that they can play the essential role of connecting psychological theory to empirical observation—the cardinal purpose of measurement. Latent variables are one such conception and network structures another. However, it would be idle to think that the possibilities are exhausted by these representations, and I hope that future psychometricians will come up with many others, so that our discipline will remain a vibrant and developing one that honors the *psycho* in psychometrics.

Acknowledgments I would like to thank Willem Heiser and Riet van Bork for valuable comments on an earlier draft of this chapter. This work was supported by NWO Vici grant no. 181.029.

References

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. & Ting, K. (1993). Confirmatory tetrad analysis. *Sociological Methodology*, *23*, 147–175.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440.

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.-M., Wysocki, A. C., van Borkulo, C. D., van Bork, R., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*(1). http://dx.doi.org/10.1038/s43586-021-00055-w.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061.

Cramer, A. O., Waldorp, L. J., Van Der Maas, H. L., & Borsboom, D. (2010). *Comorbidity: A network perspective*. https://doi.org/10.1017/S0140525X09991567.

Dalege, J. & van der Maas, H. L. J. (2020). Accurate by being noisy: A formal network model of implicit measures of attitudes. *Social Cognition*, *38*(Supplement), s26–s41. https://doi.org/10.1521/soco.2020.38.supp.s26.

de Boer, N. S., de Bruin, L. C., Geurts, J. J. G., & Glas, G. (2021). The network theory of psychiatric disorders: A critical assessment of the inclusion of environmental factors. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/article/10.3389/fpsyg.2021.623970.

Ellis, J. L. & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, *62*, 495–523.

Epskamp, S., Rhemtulla, M. T., & Borsboom, D. (2017). *Generalized network psychometrics: combining network and latent variable models*. Psychometrika. https://doi.org/10.1007/s11336-017-9557

Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2018). Network psychometrics. In P. Irwing, Hughes, D., & T. Booth (Eds.), *The wiley handbook of psychometric testing*. New York: Wiley.

Finnemann, A., Borsboom, D., Epskamp, S., & Maas, H. L. J. van der. (2021). The theoretical and statistical ising model: A practical guide in R. *Psych*, *3*(4), 594–618. https://www.mdpi.com/2624-8611/3/4/39.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, *11*, 385–398.

Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika*, *18*(4), 277–296. https://doi.org/10.1007/BF02289264.

Haig, B. D. (2005a). An abductive theory of scientific method. *Psychological Methods*, *10*(4), 371–388. https://doi.org/10.1037/1082-989X.10.4.371.

Haig, B. D. (2005b). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research*, *40*(3), 303–329. https://doi.org/10.1207/s15327906mbr4003_2.

Haslbeck, J. M. B. & Waldorp, L. J. (2018). How well do network models predict observations? On the importance of predictability in network models. *Behavior Research Methods*, *50*(2), 853–861. https://doi.org/10.3758/s13428-017-0910-x.

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

Jensen, A. R. (1999). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Junker, B. W. & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272. https://doi.org/10.1177/01466210122032064.

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., Bork, R. van, Waldorp, L. J., et al. (2018). An introduction to network psychometrics: Relating ising network models to item response theory models. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2017.1379379.

McCrae, R. R. & Costa Jr., T. J. C. P. (2008). Empirical and theoretical status of the Five-Factor Model of personality traits. In G. M. Boyle & D. Saklofske (Eds.), *G* (pp. 273–294). Los Angeles: Sage.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

Mellenbergh, G. J. (1994). Generalized Linear Item Response Theory. *Psychological Bulletin*, *115*, 300–307.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543.

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*, 461.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.

Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.

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

Reichenbach, H. (1956). *The direction of time*. Los Angeles: The University of California Press. https://doi.org/ppe.

Reise, S. P. & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*(1), 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553.

Sijtsma, K. (2006). Psychometrics in psychological research: Role model or partner in science? *Psychometrika*, *71*(3), 451–455. https://doi.org/10.1007/s11336-006-1497-9.

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

Sijtsma, K. & Molenaar, I. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks: SAGE Publications Ltd. https://doi.org/https://methods.sagepub.com/book/introduction-to-nonparametric-item-response-theory

Skinner, B. F. (1987). Whatever happened to psychology as the science of behavior? *American Psychologist*, *42*(8), 780–786. https://doi.org/10.1037/0003-066X.42.8.780.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*, 201–293.

Van Bork, R., Grasman, R. P. P. P., & Waldorp, L. J. (2018). Unidimensional factor models imply weaker partial correlations than zero-order correlations. *Psychometrika*, *83*(2), 443–452. https://doi.org/10.1007/s11336-018-9607-z.

Van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., et al. (2014). A new method for constructing networks from binary data. *Scientific Reports*, *4*(1), 5918. https://doi.org/10.1038/srep05918.

Van der Maas, H. L., 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*(2), 339–356. https://doi.org/10.1037/a0022749.