Jeanette Melin

# 5 Is validity a straightforward concept to be used in measurements in the human and social sciences?

**Abstract:** Validity in measurements in human and social sciences is commonly referred to as “measuring what one intends to measure,” and with a good fit of item parameters – somewhat simplified – it is considered to ensure validity when measuring latent traits in persons. Despite new thinking and trends about validity and positioning validity in measurement theory and practice, today’s use of validity can mostly be traced back to the classical test theory (with no compensation for ordinality, no proper separation of latent traits for persons and items, nor a defined measurement system). Consequently, when positioning models, measurement, and metrology to extend the SI, there is a need to critically review the concept of validity. A fundamental mistake is that, too often, a proper distinction is not made between the latent trait itself and the latent trait as measured. In the human and social sciences, where there are yet to be any established units, measurement-related validity should ideally not precede validity in the latent trait itself. Notably, the concept of validity has so far not been included in the International Vocabulary of Metrology (JCGM, 2012), although validation processes (entry 2.45) have been included. This is reasonably due to the centuries of work contributing to a solid consensus about the quantities in themselves. However, given the urgent needs of society for new knowledge about the world to make well-informed decisions about measurements of latent traits, we do not have centuries to first reach consensus about measurement validity. Neither was this done with the existing SI, which has been an iterative work, defining the quantities and measurement processes. Therefore, in a time where the possibilities for new units to extend the SI are being explored, an iterative and cross-disciplinary effort is needed. Thus, this chapter reviews and discusses validity and its related aspects. Finally, the chapter concludes with a proposed call for action to include a nuanced view of validity when extending models, measurement, and metrology of the SI to include measurement in the human and social sciences.

**Keywords:** validity, construct theory, construct specification equations

**Acknowledgments**: This chapter is based on learnings and outcomes in several projects during the past years, including the RISE internal platform for Categorical Based Measures and strategic initiative Measurement for Sustainable Transformation. The author would like to give a particular thanks to Leslie Pendrill and Marie-Louise Möllerberg for fruitful discussions on the topic.

**Jeanette Melin,** Division of Safety and Transport, Department of Measurement Science and Technology Unit, RISE Research Institutes of Sweden, Gothenburg, Sweden; Department of Leadership, Demand and Control, Swedish Defence University, Karlstad, Sweden

https://doi.org/10.1515/9783111036496-005

## 5.1 Why it is important to care about validity in measurements

Measuring is never an end itself; rather, it is a way of gaining knowledge about the world to make well-informed decisions. To quote Fisher (1994), *because we intend to use our measures to inform decisions that affect people's lives, we are ethically bound to be sure that the numbers represent more or less of the construct in question.* This is a general matter and not unique to measurements in the human and social sciences. However, as will be shown in this chapter, measurements in the human and social sciences face other challenges than measurements in physics, where validity is a critical component. Validity is of course important in both areas, but today for measurements in the human and social sciences, we face different challenges than in physics, and where validity is central is a critical component.

To give some examples of decisions based on measurements of latent traits in the human and social sciences, in health care, it could be questions about giving a diagnosis or prescribing treatment, setting school grades or providing support for learners with special needs, and in recruiting for work, it could be used in personnel selection (Newton & Shaw 2014). In all these cases, and many more, if the measurement does not validly capture the latent trait of interest, such decisions cannot be made validly and reliably. In health care, for instance, some patients as a consequence might be denied treatment while others who do not need it will get it, to name but one example of the serious impact of a lack of validity.

In decision-making based on measurements, it is not only validity that is important but also reliability. Figure 5.1 provides a classic picture of four dichotomous cases of measurement results to be either valid or not, and reliable or not.[1] A valid and reliable measurement is, of course, optimal in any decision-making; on the contrary, a measurement that is neither valid nor reliable is most often meaningless. In between, there is a trade-off between validity and reliability (Clifton, 2020); validity may increase with a decrease in reliability, and validity may decrease with an increase in reliability. However, the usefulness of a reliable but not valid measurement is questionable; we only know that we are measuring something well, but we do not know what we are measuring. Thus, in line with most psychometricians, we argue that validity is paramount, and reliability is contingent upon validity in measurements in the human and social sciences (Johansson et al., 2023). Furthermore, it should be emphasized that neither rigorous research design, advanced statistics, nor large samples (Flake & Fried, 2020) can make an invalid measurement valid afterward. Therefore, in the light of positioning models, measurement, and metrology to extend the SI, there is a need to critically review the concept of validity in accord with the purpose of meas-

[1] Note that this cross table also applies to the latent trait and the decision based on the latent trait as measured, which will be addressed in the forthcoming sections.

urements as a way of gaining knowledge about the world to make well-informed decisions, even those based on latent traits.

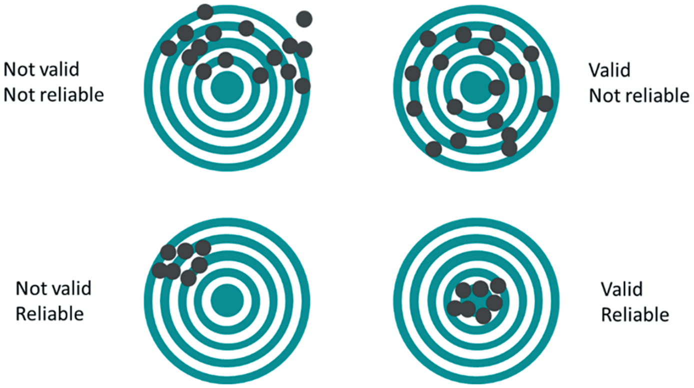


**Figure 5.1:** Four cases where a measurement can be either valid or not, as well as reliable or not. Dots that are closer to the middle indicate better validity, and more consistency of the dots indicates better reliability.

## 5.2 Latent traits and measurements of latent traits

For the traditional physical quantities and SI units (length, mass, time, etc.), for centuries, there have been internationally established agreements and definitions of the quantities themselves and procedures for measuring the quantities. However, for latent traits, there are neither such agreements about the latent traits themselves nor for the procedures of measuring the latent traits. Too often, a fundamental mistake is made when aiming to measure a specific latent trait before even understanding the existence of the latent trait itself. Therefore, we start this section by addressing latent traits themselves, followed by how to measure them.

Importantly, this chapter will not provide a discussion either on *if* latent traits exists or *if* latent traits can be measured, and such discussions, summaries, and reflections can be found elsewhere (cf. Finkelstein, 2003; Maul, Torres Irribarra & Wilson, 2016; Slaney, 2017; Michell, 2021; Mari, Wilson & Maul, 2022). Rather, when positioning models, measurement, and metrology to extend the SI into measurements in health and social sciences, our point of departure is that latent traits exist, and thus, they can be measured.

It should be noticed that this distinction between a latent trait and a latent trait *as measured* is, of course, equally important for quantities themselves and quantities as measured (Pendrill, 2019), but as will be shown later in this chapter, it has too often been forgotten, which in turn contributes to the inconsistent use of validity in measurements in human and social sciences. It is further worth to note that there is a hierarchical differentiation of quantity-related concepts and relations, as can be read in more detail in the accompanying chapter by Pendrill (2024) in this monograph. Building on the work of Dybkaer (2010) and others concerning quantities in general, but equally applicable for latent traits, concepts range from the superordinate *kind of quantities* to more specific terms, such as *quantities*, *entity quantities*, and *instantiations of an entity quantity*. Quantities as measured, as well as latent traits as measured, only have a full associative relation to instantiations of an entity quantity specifying quantity *X* itself, for entity *Y* at time *Z* to quantity *X* as measured, and for entity *Y* at time *Z* (Pendrill, 2019). However, for the readability in this chapter, we will use the shorter form: latent trait itself and latent trait as measured.

### 5.2.1 Latent traits

Latent traits are "hidden" variables, typically proposed to be within a person[2] (cf. Tesio, 2003; Battisti, Nicolini & Salini, 2010; Tesio et al., 2023) and often the main interest of the end-users. However, as will be further emphasized below, latent traits are also attributed to tasks. Examples of latent traits related to decisions in Section 5.1 could be a specific ability of a patient important for setting a diagnosis or prescribing treatment, the learner's math ability to provide grades or support for special needs, or person's attributes important in personnel selection. The basic idea is that the latent trait is unobserved but can be accessed via observed (manifest) variables conditionalizing the latent trait (Borsboom, Mellenbergh, & van Heerden, 2003). Thus, whether one observes manifest variables conditionalizing the latent trait or not, the latent trait itself can exist. Similarly, a person's mass or temperature (i.e., the quantity itself) exists independently of whether it is being measured or not, that is, at least for physical quantities there is a strong objectivity.[3]

While latent traits belonging to persons are very much what end-users are interested in the human and social sciences, nevertheless, for researchers, metrologists, psychometricians, and so on, there is another class of latent trait coupled to the latent trait of the person (Pendrill, 2014), namely latent traits attributed to tasks, which are

[2] An "agent" is the entity term superordinated to persons, organizations, cities, and so on, while patients, learners, and recruiters are subordinated to persons. The utmost work in the human and social sciences is related to persons (including subordinated groups); therefore, we use that term consistently throughout the chapter.

[3] Traditionally, a corresponding, albeit weak, objectivity is taken to apply in the human and social sciences (Pendrill 2019); a contrary position holds that the objectivity obtained both in (a) Bohm's (1952; Bohm & Hiley, 1984, 1989; Bohm, et al., 1987) ontological interpretation of quantum mechanics (Esfeld, et al., 2014; Goldstein, 1998a/b, Matarese, 2023), and in (b) Prigogine's (1971, 1976, 1978) entropy-driven theory of dissipative structures offers a potential philosophical unification of the sciences (Bohm, 2005; Bohm & Hiley, 2006; Prigogine & Stengers, 2018) exemplified by the perspective on measurement taken here (Fisher, 1988, 2024).

of significance when continuing to measurements of latent traits of persons. This is, unfortunately, unknown or neglected by many. A reason why the latent trait attributed to tasks has been less emphasized in the literature could be associated with the analogies often made between questionnaires and engineering instruments (such as thermometers). Interestingly, even though they are estimated from the margins of the same conjointly ordered data matrix, it is rarely seen that the latent trait for the items is referred to as “measured” to the same extent as measures of the latent trait for the person (see further discussion at the end of Section 5.2.2).

In fact, latent traits in themselves have nothing to do with the measurement processes. One way of illustrating this is presented in Figure 5.2, where a latent trait of a person is represented by $\theta_i$ and a latent trait of a task by $\delta_j$; both of them exist independently of each other and can be defined as $g(\chi_i, \varepsilon_i, \beta)$ and $f\left(\upsilon_j, \epsilon_j, \gamma\right)$, respectively. Thus, both latent traits have their unique sets of explanatory variables and unexplained parts. Nevertheless, how we understand and define latent traits will, of course, have implications for the measurement process, which will be addressed in the following sections of this chapter.

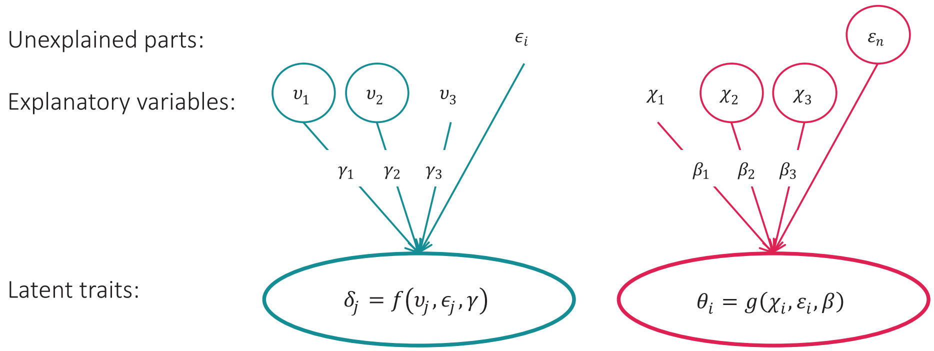


**Figure 5.2:** Illustration and notations of two coupled latent traits, for example, for tasks $\delta_j$ and persons $\theta_i$, and how they are presented as a function of both explanatory variables and unexplained parts.

## 5.2.2 Measurement of latent traits

When positioning models, measurement, and metrology to extend the SI to include latent traits, a wide definition of measurement is an important starting point (Finkelstein, 2003). Measurement can be defined as an *empirical operational procedure which assigns numbers to members of a class of entities, in such a way as to describe them; by which is meant that relations between these numbers correspond to empirical relations between the entities to which they are assigned* (Finkelstein 1975). Therefore, a critical first step toward measuring latent traits is to observe manifest tasks representing the latent trait of interest to determine how much or how little a person has of the latent trait of interest. In the simplest form, when observing manifest tasks, a person can either pass or fail, typically classified as one [1] if the test person passes or zero [0] if the test person fails. However, such classifications have no numerical meaning and only serve to indicate ordered categories (for nominal data, the categories are not ordered). Despite this well-known fact about ordinal data, counting raw scores or calculating the probability of success in a test as a measure of a test person are, unfortunately, still practiced in many fields but lack metrological quality assurance.

The basic idea of observing manifest tasks representing the latent trait of interest is very similar when advancing the methods to ensure measurement quality. In fact, as has been noted for some decades (Andrich, 1978, 1988, p. 43; Engelhard, 2012; Linacre, 1995, 2000a/b; Wright, 1997), multiple independent developments (Bradley & Terry, 1952; Luce, 1959; Luce & Tukey, 1964; Peirce, 1878; Rasch, 1960; Thurstone, 1928; Zermelo, 1929) show that ordinal observations can be restituted into interval measurements via models defining unit quantities that retain their properties independent of the questions asked and persons responding to within a fit-for-purpose degree of uncertainty. Thus, there are two critical phases for providing measurements of latent traits:

- the observation phase, that is, when data is collected, for instance with a questionnaire, observation protocol, or test from a person or a group of persons, and
- the restitution phase, that is, when separating the probability of success from the observed data into separate measures of the two latent traits (attributed to persons and tasks).

When considering the basic assumption of measuring latent traits – that a person who has more of the latent trait will be more likely to score higher on a difficult item (i.e., manifest task) than a person who has less of the latent trait, and conversely, it is more likely that more persons score high on an easy item – the importance of quantifying both latent traits and their relationship might become clearer. This relation between the two coupled attributes is given by the formula (Rasch 1960; Wright & Stone, 1979):

$$P\left(\pi_{ij} = 1 | \theta_i, \delta_j\right) = \frac{e^{\left(\theta_i - \delta_j\right)}}{1 + e^{\left(\theta_i - \delta_j\right)}} \tag{5.1}$$

where the probability $\pi$ of a response scored 1 from person $i$ in relation to task $j$ is a function of the difference between $\theta_i$, the latent trait attributed to the person, and $\delta_j$, the latent trait attributed to the tasks. Rasch's (1960, pp. 110–115) formulation of this model originated in an analogy from Maxwell's treatment of Newton's second law of motion, but most early applications were in the educational sciences, where the common association of latent traits attributed to persons are typically referred to as abilities and latent traits attributed to tasks are typically referred to as difficulties. In this chapter, when we provide examples, for simplicity, we will use person's ability and task's difficulty, but the thinking, of course, can also be extended to other latent traits.

Figure 5.3 links the latent traits themselves (Figure 5.2) via the observation phase and the measurand restitution (eq. (5.1)) to the latent traits as measured (for items $\delta_{j,m}$ and persons $\theta_{i,m}$). At the top of the figure, we have two latent traits that need to be "coupled," and the observed response depends on both the latent trait attributed to the person and the latent trait attributed to the task. As we often intend to measure the abilities of persons, we can also refer to the observation phase where the observed response will depend on the person's ability and the item's difficulty. In the next step, with measurand restitution – here done by estimating the parameters in eq. (5.1) – separate measurements of the coupled latent traits can be obtained. Thus, we are using the manifest tasks to provide measurements of the latent trait of persons, and we are using persons to provide measurements of the latent trait of the tasks.

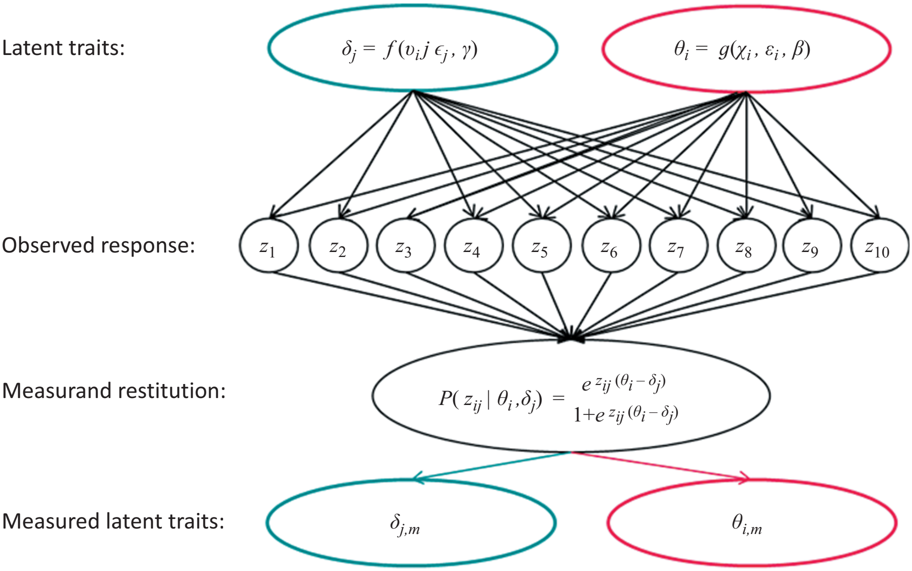

Latent traits:
$\delta_j = f(\upsilon_i j\, \epsilon_j, \gamma)$
$\theta_i = g(\chi_i, \varepsilon_i, \beta)$
Observed response:
$z_1$ $z_2$ $z_3$ $z_4$ $z_5$ $z_6$ $z_7$ $z_8$ $z_9$ $z_{10}$
Measurand restitution:
$P(z_{ij} \mid \theta_i, \delta_j) = \frac{e^{z_{ij}(\theta_i - \delta_j)}}{1+e^{z_{ij}(\theta_i - \delta_j)}}$
Measured latent traits:
$\delta_{j,m}$
$\theta_{i,m}$

**Figure 5.3:** Illustration and notations for how the latent traits themselves (Figure 5.2) are coupled into the observation phase and through the restitution phase provide separate measures of latent traits, for example, tasks $\delta_{j,m}$ and persons $\theta_{i,m}$.

The observed response does not, however, depend only on the latent trait of the person and the latent trait of the task. There are, in fact, additionally other components from the measurement process that are not yet fully compensated for in the model shown in eq. (5.1). For example, Figure 5.4 shows a more complete picture for latent traits of the measurement process (Pendrill, 2019; Pendrill, 2014; Bentley, 2005; Pendrill, 2023), where measurement information is transmitted from the measurement *object,* via an *instrument* to an *operator* in the observation phase, which both the *environment* and the *measurement method* can influence.

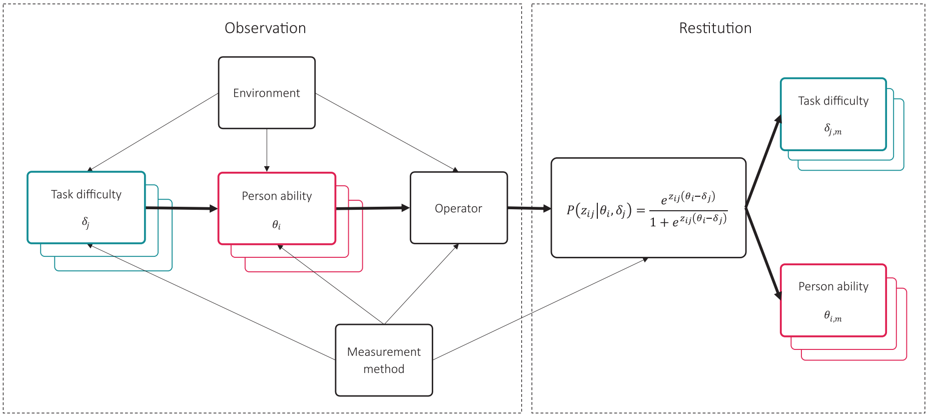


**Figure 5.4:** An illustration of the measurement system for latent traits linking the observation phase with the restitution phase. Tasks, for example, questionnaire items, provide stimuli due to their difficulty to the test persons who respond to each item, where the response depends on both the difficulty of the task and the ability of the person (i.e., the latent trait themselves, for items $\delta_j$ and person $\theta_i$), which in turn can be restituted with the model shown in eq. (5.1) into measurements of tasks $\delta_{j,m}$ and persons $\theta_{i,m}$.

This view of the measurement system corresponds more directly with the traditional approach in engineering science and technology than with typical arguments in human and social sciences measurements. Much is to be gained by adopting this approach. Specifically, in measurement engineering (Bentley, 2005), an instrument converts an input (such as from a stimulus from the measurement object) into an output response, while the measurement object has no input but only produces an output (which acts as a stimulus input to the instrument), for example from weighing, where an object has a mass that stimulates the instrument (scales) to respond with an indication of the mass. Similarly, in both traditional and “psychometric” measurement systems, the measurement object (weight or task) is a natural first choice of metrological standard – with its robustness and simplicity – in preference to the relatively sensitive and complex instrument, with more sensitivity to the environment, context, and method (Pendrill, 2021; Melin, 2021). Thus, Pendrill (2018) has argued that: *drawing simple analogies between “instruments” in the social sciences questionnaires, ability tests, etc. – and engineering instruments such as thermometers does not go far enough*. As will be shown later in this chapter, a complete picture of the measurement process and the measurement system will have implications for using the concept of validity, which is significant when positioning models, measurement, and metrology to extend the SI.

Notably, in contrast to measurements in physics, calibration and the measurement itself are often done simultaneously for measurements in human and social sciences. For example, while arguing that a set of items, that is, an item-bank, is analogous to a calibrated set of weights to ensure metrological comparability when measuring person’s ability (Pendrill, 2018), previously existing measurements of task’s difficulty are not always being used for measuring person's ability in a new cohort. Nevertheless, with the model shown in eq. (5.1) (Rasch 1960), measurements of person's ability are easily restituted based on the raw scores from the observation phase based on previously existing measurements of task's difficulty. In turn, this will enable comparability beyond the present cohort of persons. Another way, even more accessible, to achieve measurements of person's ability is, thanks to conversion tables, again where raw scores from the observation phase are being used and converted to measures in the same way that meters can be converted to inches (Melin et al., 2023a).

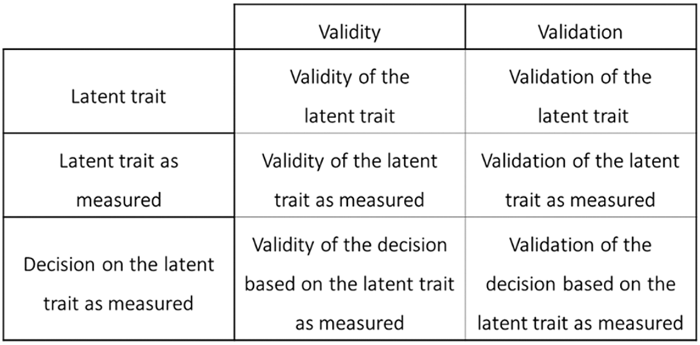
## 5.3 Validity and latent traits

History shows that validity concepts in measurements in human and social sciences has undergone a "metamorphosis" (Geisinger 1992). Although validity in measurements in human and social sciences is commonly referred to as *whether a test measures what it purports to measure* (Kelley 1927), it is only sometimes reflected in practice when choosing theories and methods. Many others have made good summaries of the evolvement of validity as a concept in the human and social sciences (cf. Messick 1989a; Newton & Shaw, 2014; Borsboom, 2005; Slaney, 2017; Kane, 2016), and such summaries go beyond the scope of this chapter. However, we will instead pick up some of the key contributions to today's somewhat fragmented use of validity and will, at the end of this section, return to and review the statement by Borsboom et al. (2004), claiming that *validity is not complex, faceted, or dependent on nomological networks and social consequences of testing.*

### 5.3.1 Validity and validation

The first very fundamental differentiation is between validity and validation: validity is about ontology, and validation is about epistemology (Borsboom, Mellenbergh, & van Heerden, 2004). First, we argue for the need to consider validity aspects related to both the latent traits themselves (Section 5.2.1) and the latent traits as measured (Section 5.2.2), that is, the trueness of both the existence of the latent trait and of the measurement results. Furthermore, Wolf et al. (2019) summarized the contemporary validity literature as saying that *validity is not an inherent feature of a survey (or other instruments) but rather a characteristic of the survey concerning a particular use* [. . .] *as a consequence, validation is necessarily fit-for-purpose, such that different forms of argumentation and evidence may be necessary depending on the design and intended purposes of the survey.* This gives us three potential validity claims: the validity of the latent trait, the validity of the latent trait as measured, and the validity of the decision based on the latent trait as measured.

The literature has been moving away from the concept of validity and emphasizing instead methods for validation, which are – at least theoretically – also applicable to all three validity claims. Figure 5.5 summarizes those distinctions for any latent trait. However, below, it will be put in the context of the two coupled attributes: a latent trait attributed to persons and a latent trait attributed to tasks. In addition, it should be noted that there is a further question: Is the validation valid? That is, one must distinguish between two kinds of decisions: the validity of the decision on the latent trait as measured and the validity in the claim about the validation of the latent trait.



| | Validity | Validation |
|---|---|---|
| Latent trait | Validity of the latent trait | Validation of the latent trait |
| Latent trait as measured | Validity of the latent trait as measured | Validation of the latent trait as measured |
| Decision on the latent trait as measured | Validity of the decision based on the latent trait as measured | Validation of the decision based on the latent trait as measured |

**Figure 5.5:** The distinction between validity aspects and validation for the latent trait, the latent trait as measured, and the decision based on the latent trait as measured.

When extending models, measurement, and metrology of the SI into measurements which also cover the human and social sciences, of course, one needs to consider the International Vocabulary of Metrology (VIM) (JCGM, 200:2012). Notably, validity is not yet included in the VIM, while validation (entry 2.45) is. This reasonably is due to the centuries of work to reach a consensus on the physical quantities[4] in themselves. However, an important note is that validation is defined as: *verification, where the specified requirements are adequate for intended use*, reflecting validation of the measurement process rather than validation of the quantities themselves or quantities as measured.

To summarize this section, validity and validation are distinct concepts that should not be mixed. One must be careful when making claims about measurement-related validity and decision-related validity before the validity in the latent trait itself has been ensured. In a time where the possibilities to extend the SI for new units are being explored, it is, however, important to stay open for an iterative and cross-disciplinary effort to advance both the validity of the latent traits themselves and the measured la-

[4] Since 1968, within the SI units, there are not only physical quantities but also mol for the amount of substance.

tent traits as well as developing methods for validation of latent traits themselves and the measured latent traits. This will be further discussed in Section 5.4.

### 5.3.2 The many facets of validity

A decade ago, Newton and Shaw (2014) published a book about validity in educational and psychological testing, including a list of 151 (!) different kinds of validity. Based on decades of research, they summarized three different claims related to validity:

1. Validity as a **measurement claim**: It is possible to measure a latent trait accurately using a measure of the latent trait.
2. Validity as a **measurement and decision-making claim**: It is possible to make accurate decisions on the basis of measurements of the latent trait.
3. Validity as a **concept spanning measurement, decision-making, and broader impacts and side-effects**: It is acceptable to implement a measurement policy.

While none of these actually address the validity of the latent trait itself, that is, if the latent trait exists or not, the first claim relates very much to the original statement of validity – *whether a test measures what it purports to measure* (Kelley 1927). This is also related to the “middle validity claim” in Figure 5.5 (i.e., the validity of the latent trait as measured). The second and third validity claims are reasonably a response to the significance of being able to justify interpretations and actions concerning social and ethical consequences of test use (Messick 1989a, 1989b) and the separation of different kinds of validity (Joint Committee on the Standards for Educational and Psychological Testing, 2014; Cronbach & Meehl 1955). Tracing back to the mid-nineteenth century, three types of validity dominated, namely content, construct, and criterion validity. Traditionally, both content and construct validity relate to how a set of test items can be used to measure a person’s latent trait of interest validly; content validity is typically referred to if the set of test items reflects the important components related to a given person’s latent trait and construct validity on the psychometric properties of the used set of items. On the other hand, criterion validity is more related to how measurement values of the person’s latent trait can either be compared with results from similar measurements (also known as concurrent validity) or predict an outcome at a later time (also known as predictive validity). Those three types of validity have different significance in different contexts, where content validity has a particular role in achievement tests, construct validity for personality tests, and criterion validity for an aptitude test (Newton & Shaw, 2014). We acknowledge this tradition and understand that different aspects may have different importance for the end-user, but simultaneously believe that this confuses the use of the validity term.

The use of criterion-related and, in particular, predictive validity has been and continues to be dominating in personal selection. For this purpose, the persons’ measurement values of the latent trait are viewed as a *sign or signal of future performance* *and rely on evidence that individuals with higher predictor scores* [i.e., measurement value] *subsequently perform better* (Van Iddekinge, Lievens, & Sackett, 2023). Thus, the main focus is on the measurement value and its relation to the future rather than what the measurement value stands for. Roughly speaking, if measurement values from persons based on a set of items can predict future outcomes well, then the prediction is more important than the latent trait itself and the latent trait as measured. Here, greater "allowances" to focusing on reliability at the expense of validity are often accepted (Clifton, 2020).

Since the 1950s, the American Psychological Association, the American Educational Research Association, and the National Council on Measurements in Education have been leading actors in the field of validity, including publishing of the *Standards for Educational and Psychological testing*. Initially, focus was on the three parts of validity (i.e., content, construct, and criterion validity), but in the third (and fourth) edition, there has been a shift toward considering validity as multidimensional and complex, requiring a wide and diverse body of evidence (Goodwin & Leech, 2003). The *Standards* comprise the following validity-related concepts (Joint Committee on the Standards for Educational and Psychological Testing, 2014):

- Evidence-based **test content** refers to the set of items that represents the domain it proposes to measure (similar to content validity)
- Evidence-based **response processes** are the extent to which different types of respondents' responses fit the defined construct (similar to construct validity)
- Evidence-based **internal structure** is about how the components match the defined construct (similar to construct validity)
- Evidence-based **relations to other variables** reflect expected relations based on the theory of the construct being assessed (similar to criterion validity)
- Evidence for **validity and consequences of testing** includes both anticipated and unanticipated consequences of the measurement.

Building on McAllister's (2008) claim that probabilistic conjoint measurement offers *a statistical model for validating assessment tools that are particularly suited to quantifying human performances on assessment items*, Mui Lim et al. (2009) proposed examples of validation activities and validation linked to the types of validity in the *Standards* (Joint Committee on the Standards for Educational and Psychological Testing, 2014). While we would, in line with Pendrill (2014), stress that identified measurement models' (San Martin & Rolin, 2013) testing for conjoint additivity (Newby & Bunderson, 2009), parameter separation, and specific objectivity (Rasch, 1966) are *not simply a mathematical or statistical approach but instead a specifically metrological approach to human-based measurement*, the proposal of how Mui Lim et al. (2009) suggests validation activities are very welcomed in relation to the view of validity provided by the *Standards*. On the contrary, in view of the weakness in the *Standards* of not thoroughly addressing validity in the latent trait itself, such validation activities can be carefully implemented, provided one has firstly ensured the validity of the latent trait itself.

Furthermore, it has been proposed that the statistical sufficiency principles of measurement modeling (i.e., that the observed score should capture all available information in the data; Andersen, 1999; Andrich, 2010) are related to Messick's construct validity issues (Baghaei, 2008), namely, *that nothing important be left out* (Messick 1996; Messick 1994). Here, Baghaei (2008) argues for assessing model fit statistics to indicate possible construct-irrelevant variance and assessing the conjoint item-person histogram to assess construct underrepresentation. He further links some different types of validity to typical evaluations as to whether the items address the intended latent trait, the item difficulty hierarchy makes sense as an expression of the construct, and measurement values correlate well with measurements estimated from other sets of items probing the same latent variable. On the other hand, questions as to whether the person's ability hierarchy makes sense indicate that this approach to predictive validity may be less related to the common practice of comparing measurements to an outcome estimated via other means at a later time.

This section has provided a short summary of the many facets of validity, including some work specifically related to probabilistic conjoint measurement (as it is key in the measurement restitution process). Work, so far, has been dominated by the latent trait as measured for persons. This can reasonably be explained by end-user interest in making decisions about the persons based on measurement. It is also likely that the classical test theory – that is, where no proper separation is made between the latent traits of persons and items are being made – has impacted the setting of the terminology and use of it. Nevertheless, we agree with Stenner (2014) that *validity should be equally applicable to both latent traits* in measurements in the human and social sciences.

### 5.3.3 Revisiting the bull's eye target

It appears that Figure 5.1 – the classic bull's eye target for illustrating validity and reliability – could be revisited by asking what target is aimed for the closeness to the latent trait itself or the closeness to the measurement of the latent trait based on a reference method? Figure 5.6 highlights four possible different targets for the bull's eye. Obviously, the closeness to the latent trait itself relates to the overall aim of measuring, that is, a way of gaining knowledge about the world to make well-informed decisions. At the same time, comparing a measured value of a latent trait with a value of the latent trait itself is impossible because the value of the latent trait itself is not accessible. Despite that, it is likely that much of the literature refers to the middle point as the latent trait itself when not properly distinguishing between the latent trait itself and the latent trait as measured.

In an illustrative case (Figure 5.7), imagine analyses of two questionnaires claiming to measure the same latent trait of persons but with different sets of items. Suppose that both fit the measurement model well (e.g., no significantly misfitting items, no differential item functioning, no local dependency beyond the modeled stochasticity of the

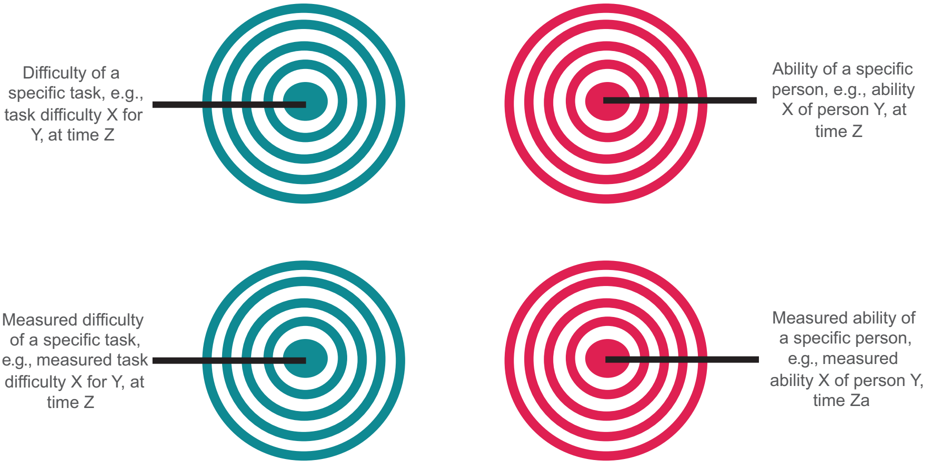

Difficulty of a specific task, e.g., task difficulty X for Y, at time Z
Ability of a specific person, e.g., ability X of person Y, at time Z
Measured difficulty of a specific task, e.g., measured task difficulty X for Y, at time Z
Measured ability of a specific person, e.g., measured ability X of person Y, time Za

**Figure 5.6:** Four cases with different possible targets in the bull's eye. To the left, targets are related to task's difficulty and to the right targets are related to person's ability. The upper boards are related to the latent trait themselves, and the lower boards are related to the latent traits as measured.

data, unidimensionality usefully approximated to within the tolerance limits of the application, and response categories that work as intended (Johansson et al., 2023)), but when correlating to another variable, the association is different. This raises questions such as how to make a valid inference about the association between *y* and the latent trait and which one is the "optimal" way of gaining knowledge about the world to make well-informed decisions? In fact, similar results have empirically been shown by Maul (2017), for instance, by including two set of items intended to measure growth mindsets with the notable exception that the key noun in the sentence ("intelligence") had been replaced with a nonsense word ("gavagai") in one of the item sets. Analyses of both sets of items, however, fitted models well. To quote Maul (2017), *it would seem difficult to take seriously the claim that any of these sets of items constituted a valid measure of a psychological attribute, and if such a claim were made, one might reasonably expect any quality-control procedure worthy of the name to provide an unequivocal rejection*. Thus, in cases where interpretation of the bull's eye becomes a measurement issue, it cannot be separated from a qualitative understanding of the latent trait itself.

### 5.3.4 Designing measurements of latent traits

As shown in Figure 5.2, a latent trait of a person, $\theta_i$, and a latent trait of a task, $\delta_j$, can exist independently of each other, while they show a special relationship when measuring latent traits (eq. (5.1)). To respond to the most common end-user need – measurement values of a specific latent trait attributed to persons – it is natural to start by defining the latent trait related to the person and after that designing items to be

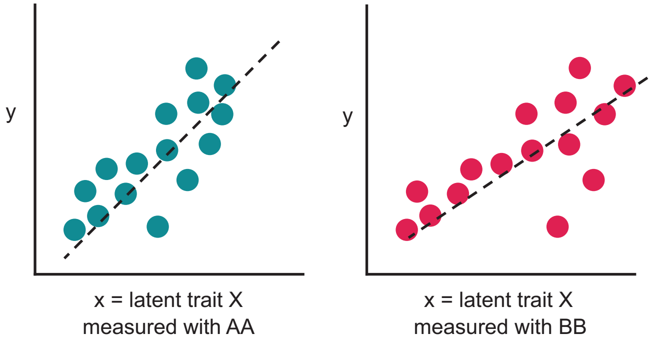


**Figure 5.7:** A fictive illustration for the same latent trait, *x*, measured with two different set of items that both fit the model well, correlated to *y*.

used to assess what it means for persons going up or down the scale (Wilson, 2005). However, historically, some psychologists have tended to view what is being measured as an empirical matter with a conception that views validity as something to be discovered afterward (Borsboom, Mellenbergh, & van Heerden, 2004). Furthermore, Fisher (1994) rhetorically asked *validity by default or design* and continued to claim that *just because experts have decided that items on a test all belong to the same content domain does not mean that they belong to the same construct*. Therefore, although having experts provide their views on what it means to go up or down the scale is a critical starting point, it is not enough to claim validity in either the latent trait itself or the latent trait as measured. Likewise, empirical data fitting the measurement model may help understand the latent trait itself and design measurements. Again, it is not enough in itself to claim validity in either the latent trait itself or the latent trait as measured.

As a key aspect when measuring latent traits, Morel and Cano (2017) stressed that *of all measurement properties, "content validity" is a sine qua non*. Indeed, rigorous research design, advanced statistics, nor large samples (Flake & Fried, 2020) can compensate for this afterward. A proper design to overcome limited – or, at worst misleading – measurements of a latent trait includes a *substantive patient-driven clinically anchored framework* (Morel & Cano, 2017), extending beyond health care in the human and social sciences. There is, however, an important differentiation between what is being used to measure a specific latent trait of interest (e.g., which items in a questionnaire) and what latent traits are of interest when making decisions. For the first point, setting up a set of items to be used to measure a specific latent trait attributed to the persons, must be carried out as a noncompetitive activity combining different expertise and resources (Morel & Cano, 2017). While, for instance, patients are key partners when developing measurements in health care, metrologists with expertise in latent traits must also be viewed as key partners. For example, they have a unique expertise in what requirements are important for designing good measurements. Therefore, they should be able to facilitate the work to hypothesize the composition of a latent trait of interest and how item can be mapped hierarchically along a clinical continuum (Barbic, Cano, & Mathias, 2018) and evaluate how well empirical data fit the measurement model as one source of information when developing measurements of latent traits. Designing items need to ensure enough variation in the item contents while at the same time staying within one dimension, and by asking enough questions to reduce uncertainty in relation to that variation (Fisher, Melin, & Möller, 2021). Furthermore, a good example of the latter comes from Morel and colleagues (2022), who provided a conceptual model for experiences from the early stage of Parkinson's disease, and in turn, lay the foundation of what latent trait to be measured in order to make a decision based on what matters for that target group.

The literature about designing measurements for latent traits, again naturally, starts by defining the latent trait related to the person and, after that, designing items (Wilson, 2005). A subsequent step when intending to test and evaluate the measurement in research is the study design. Much of that is, however, also a part of the description of the full measurement system, equally important to be considered and optimized in all measurement situations to give as good as possible measurements of the coupled latent traits. Recalling Figure 5.4, which shows a complete picture of the measurement process (Pendrill, 2019; Pendrill, 2014; Bentley, 2005) for latent traits, measurement information is transmitted from the measurement *object,* via an *instrument*, to an *operator* in the observation phase, which both the *environment* and the *measurement method* can influence. Design aspects of the operator, environment, and method also become apparent. Table 5.1 presents the measurement system components, entities for latent traits in general, and an example with memory measurements (Melin & Pendrill, 2022a).

Koopmans (1947) contrasts Tycho Brahe and Johannes Kepler's work nicely, where Brahe took a systematic approach of careful measurements, while Kepler looked for new models and was able to find simple empirical "laws" which were in accord with past observations as well as permitting the prediction of future observations. Combining theory-driven designs with an open-minded, explorative approach in an iterative and cross-disciplinary environment may foster the curiosity needed when new units to extend the SI are being explored. For instance, items – or persons – that do not fit the model may indicate multidimensionality and candidates for modification or discarding a theory (Baghaei, 2008). Likewise, items that do not match the expected hierarchy from theory or previous studies may warrant theory refinements (Karlsson et al., 2023). In particular, such anomalies will guide when and where to look for new phenomena (Kuhn, 1977) and pieces in the understanding of both the latent trait itself and the latent trait as measured. In turn, this need for iterative work will present possibilities for the design of measurements in the human and social sciences (Fisher & Stenner, 2011).

**Table 5.1:** Entities in the measurement system in general, for latent traits in general and exemplified for memory measurements.

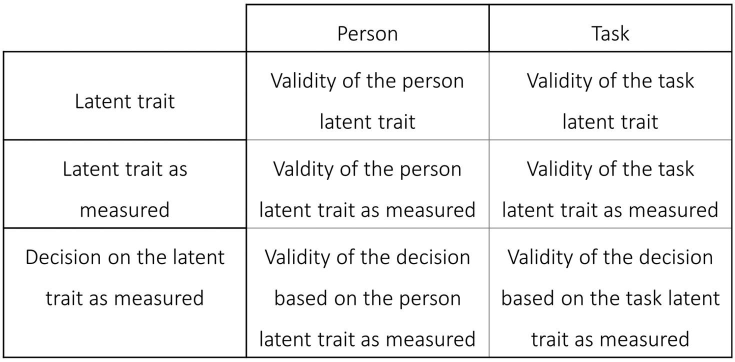
| MSA term | Entities for latent traits in generall | Entities for example for memory measurements |
|---|---|---|
| Object | One test item | A sequence of numbers to be recalled |
| Objects | A set of test items | Sequences of numbers to be recalled |
| Instrument | One person | One person whose memory is being measured |
| Instruments | A cohort of persons | A cohort of persons whose memories are being measured |
| Operator, example role 1 | The test leader | Study nurse |
| Operator, example role 2 | The person observing | Study nurse |
| Operator, example role 3 | The person doing the measurand restitution | Person doing the Rasch-analysis |
| Environment | The context of the measurement | Time on the day, place for testing |
| Method | Specifications in the observation phase | Item order, presentation of items |
| | Specifications in the measurand restitution phase | The measurand restitution |

### 5.3.5 Construct specification equations as a mean of validity

Despite the fact that observed data can be provisionally and initially validated to a limited degree simply via fit to a probabilistic conjoint measurement model (eq. (5.1)), a measurement that lacks a construct theory is, as stated by Stenner et al. (2013), *a black box in which understanding may be more illusory than not*. Thus, more is needed to claim validity than just the fit of the data to a model and the demonstration of group invariance. In line with that, Stenner and colleagues (1982; 1983) introduced the so-called construct specification equations (CSEs) for latent traits attributed to tasks, which along with related approaches to devising explanatory measurement models (De Boeck & Wilson, 2004; Embretson, 2010; Fischer, 1973) are frequently stressed as a mean of validity when measuring latent traits of persons (McKenna et al., 2019; Stenner et al., 2006; Stenner et al., 2013; McKenna, Heaney, & Wilburn, 2019). Specifically, the CSE approach has to date been used mostly to explain the latent trait itself for tasks as an argument for the validity of the measured person attributes:

> *The rationale for giving more attention to variation in item scores is straightforward. Just as a person scoring higher than another person on a set of items is assumed to possess more of the construct in question (i.e., visual memory, reading comprehension, anxiety), an item (or task) that scores higher (in difficulty) than another item must be viewed as demanding more of the construct. The key question deals with the nature of the "something" that causes some persons to score higher than others and some items to score higher than others.* [. . .] *Such an equation embodies a theory of item-score variation. It simultaneously provides a straightforward means of confirming or falsifying alternative theories about the meaning of scores generated by a measurement procedure.* (Stenner & Smith 1982)

While Stenner and colleagues made a substantial contribution to advancing methods for validation – including CSE – of measurements in the human and social sciences, statements such as *an instrument is valid if it measures the intended attribute*, and *the "validity" concept should be equally applicable, to both attributes* (Stenner, 2014) is, however, somewhat contradicting. It may, to some extent, be explained by the lack of separating the latent trait itself and the latent trait as measured, as well as an improper description of the measurement system. Combining the fact that validity is applicable for both latent traits as well as for both the latent trait itself and the latent trait as measured is summarized in the matrix in Figure 5.8. Even though the main interest of the end-user is most often associated with making decisions about a latent trait attributed to the person, decisions attributed to the tasks are also applicable.[5]



| | Person | Task |
|---|---|---|
| Latent trait | Validity of the person latent trait | Validity of the task latent trait |
| Latent trait as measured | Valdity of the person latent trait as measured | Validity of the task latent trait as measured |
| Decision on the latent trait as measured | Validity of the decision based on the person latent trait as measured | Validity of the decision based on the task latent trait as measured |

**Figure 5.8:** The distinction of validity claims for the latent trait, the latent trait as measured, and the decision based on the latent trait as measured separated for latent traits attributed to persons and tasks.

We would argue for the CSE itself to be a "model" of the latent trait itself and consequently as a means of validity. In turn, the validity of the measured latent trait can be obtained when there is a high correlation between the measured latent trait and the latent trait itself, which applies for both coupling attributes such as person ability, $\theta$, and

[5] An example of where a decision about the task(s) is the main interest is psychophysics, where a test panel is used to quantify specific latent traits attributed to products.

task difficulty, $\delta$. In line with Stenner and colleague's work, we have also initially focused on CSEs for task difficulty (Melin et al., 2019, 2022a, 2022b; Pendrill, 2019; Melin, Cano, & Pendrill, 2021; Melin & Pendrill, 2023), although not only as a means for the validity of the measured latent trait of the person. Rather, the point of departure for CSEs for task difficulty has mainly been driven by the measurement system approach (presented in Section 2.2) where the human responder acts as the instrument (Pendrill, 2014). Specifically, we have suggested that CSEs *appear to provide metrological references for calibration and subsequent inter-comparability of measurements* (Melin et al., 2022b). Particularly, CSE can not only serve as a means of validity but also resemble formulas for "reference measurement procedures" (RMPs) analogous to RMPs found in the metrology of chemistry.

While the pure theoretical definitions of a latent trait attributed to a person $\theta_i$ can be defined as $g(\chi_i, \varepsilon_i, \beta)$ and a latent trait attributed to a task $\delta_j$ can be defined as $f\left(\upsilon_j, \epsilon_j, \gamma\right)$ (Figure 5.2), the CSE, however, can be considered a quasi-theoretical model of the latent trait itself:

$$\hat{\mathbf{Z}} = \sum_k \beta_k \cdot x_k \tag{5.2}$$

where $\boldsymbol{Z}$ is the latent trait of interest. In turn, $\mathbf{Z}$ is defined as a linear combination of a set, $\boldsymbol{k}$, (independent) variables, $\boldsymbol{X}$. Similar to a purely theoretical model, the CSE is equally applicable to both latent traits (Figure 5.9). Thus, some variables that cause variation in the latent trait attributed to persons explain why some people have better abilities than others. Likewise, some variables that cause variation in the latent trait attributed to tasks explain why some tasks are easier than others.

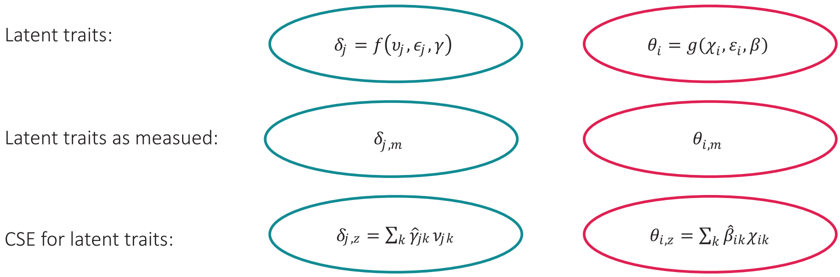

Latent traits:
$\delta_j = f\left(\upsilon_j, \epsilon_j, \gamma\right)$
$\theta_i = g(\chi_i, \varepsilon_i, \beta)$
Latent traits as measued:
$\delta_{j,m}$
$\theta_{i,m}$
CSE for latent traits:
$\delta_{j,z} = \sum_k \hat{\gamma}_{jk} \nu_{jk}$
$\theta_{i,z} = \sum_k \hat{\beta}_{ik} \chi_{ik}$

**Figure 5.9:** Notations separated for latent traits themselves, latent traits as measured, and CSE for latent trait separated for latent traits attributed to tasks $\delta_j$ and persons $\theta_{i.}$

In the EMPIR projects, NeuroMET 15HLT04 and NeuroMET2 18HLT09, researchers from national metrology institutes, academia, and industry have worked together to improve measurements for neurodegenerative diseases (Quaglia et al., 2021). One work package has been dedicated to memory measurements, which is one of the first metrological projects in European level to include measurements of latent traits. In the development of the NeuroMET Memory Metric, CSEs have been used as means of validity claims when combining different items from legacy tests. Block and digit recalling items reveal almost identical CSEs (Melin, Cano, & Pendrill, 2021), and two kinds of word recalling items reveal almost identical CSEs (Melin et al., 2022a; Melin & Pendrill, 2022). Furthermore, with entropy – originated from the Brillouin expression – dominating all CSEs (Melin et al., 2022b), they add validity that goes beyond a good fit to a measurement model (Melin et al., 2023a). Even though corresponding CSEs – including the dominating entropy contribution – for backward recalling block and number tasks have also been studied (Melin et al., 2023b), in the NeuroMET Memory Metric, only forward recalling sequences are included. This is because indications were found of multidimensionality when combining forward and backward recalling sequences as well as the set of items challenging the test person in terms of maintenance or manipulation working memory, respectively, and the constructs appeared less related and more likely to represent different underpinning constructs (Melin et al., 2023b). Thus, as argued above, fit statistics and a qualitative understanding are important, and this needs to go hand in hand also with the CSE.

An important note is that our CSE approach differs from the earlier work by Stenner and colleagues (1982; 1983) in choosing a principal component regression (PCR) rather than a regression based on the explanatory variables. This is because we cannot be sure *how* independent the explanatory variables are and whether they are the experimentally observed quantities or not (Melin & Pendrill, 2023). Specifically, when applying a principal component analysis in the PCR, one identifies the main components of variation by "rotating" in the explanatory variable space from the experimental dimensions to the principal component dimensions. Thus, when using principal components, we can allow some combination of the explanatory variables in cases where there is a significant correlation between them. A second important note is why we consider the CSE to be quasi-theoretical. This is because the linear regression is being made of the latent traits as measured – for persons $\theta_{i,m}$ or tasks $\delta_{j,m}$ – against $\boldsymbol{X}'$ in terms of the principal components. Ideally, we would have made the regression of the latent trait themselves, but obviously, it is not accessible. Thus, the measurement values of the latent trait are the closest to being used. It should, however, be noted that it warrants a good qualitative understanding of what is being measured and a good fit to the model to avoid developing misleading CSEs. On the other hand, a CSE may not only serve as a means of achieving validity, but as will be discussed in Section 5.4, it may also be used as an explorative tool when advancing the understanding of both the latent trait itself and the latent trait as measured when positioning models, measurement, and metrology to extend the SI.

### 5.3.6 Is validity straightforward or complex?

To close this section, we pick up on Borsboom et al. (2004), who claimed that *validity is not complex, faceted, or dependent on nomological networks and the social consequences of testing.* We agree that the concept's meaning can be very straightforward; nevertheless, the use of it has not been straightforward. Consequently, while validity has multiple meanings in measurement in the human and social sciences, a first step toward a more unified view of validity must separate the three validity claims presented in Figure 5.5, where the validity of the latent trait itself is often (or perhaps always?) a precondition to the validity of the latent trait as measured, which is often (or perhaps always?) in turn a precondition for the validity of the decision based on the latent trait as measures.

## 5.4 Routes to a better use of validity terminology and processes when extending the SI

Despite the fragmented use of validation processes, all agree that validity should be optimized. This work deals with it indirectly, but our key message calls for a better – optimized and clearer – terminology for validity. In turn, we believe that it will advance the validity of the latent traits themselves, the validity when measuring latent traits, and the validity in decisions about latent traits. Thus, the most important message is to understand the difference between the latent trait itself (Section 5.2.1) and the latent trait as measured (Section 5,2.2), and consequently, three important claims of validity need to be distinguished (Figure 5.5).

Furthermore, in a time when both latent traits ought to be understood, ways to measure the latent traits, and finding methods for validation are needed, we would encourage an iterative and cross-disciplinary approach rather than a too strict process. This is expected to advance the field of measurements in the human and social sciences when extending models, measurement, and metrology of the SI. At the same time, one must be careful not to make too strong validity claims.

### 5.4.1 Iterative, explorative, and cross-disciplinary efforts when measuring latent traits

On the one hand, clearly and consensus-based "rules" for validity in the human and social sciences when extending models, measurement, and metrology of the SI will support a more "production-like" process. On the other hand, it might be that the field is not yet ready for a "one-size-fits-all" approach. Of course, again, there is a need to have a harmonized view of terminology and possible limitations in claims at different stages. For instance, a too-hardline data-driven approach could be dangerous (Morel & Cano, 2017). Even when items are designed with a construct theory in mind, it might happen that observations do not vary as expected or do not fit the measurement model. For instance, uninterpretable inconsistencies might be due to an underdeveloped theory and/or low-quality data (Fisher, Melin, & Möller, 2021), but this should not be *a sign of the end of the conversation or of the measurement effort* (Fisher & Stenner, 2011).

Rewriting and/or changing items is the most common way of addressing the issue with the misfit. However, other aspects might also be related to the measurement system and the measurement process affecting model fit to be considered, adjusted for, and re-evaluated. For instance, how did the test leader interact with the test person, when and where was the observation, and what kind of specifications were used in the measurement restitution (Table 5.1)? As an example of the latter, specific objectivity (Rasch, 1966) is a unique feature of probabilistic conjoint models requiring separable parameters and minimally sufficient statistics, implying that the comparability of measurements of latent traits attributed to the person should be independent of which test items are being used and, symmetrically, comparability of item measures should be independent of which test persons are being used. In contrast with this capacity to support metrological traceability, other classes of models, such as those falling under the heading of "Item Response Theory" (IRT; Hambleton et al., 1991), cannot maintain unique metrological properties (Embretson, 1996; San Martin et al., 2009, 2015).

We do not assert the metrological viability of sociocognitive measurement without recognizing that

– local realizations and interpretations of even physical units of measurement may vary across communities of research and practice in divergent ways (Galison, 1997; Tal, 2014; Woolley & Fuchs, 2011);
– that irreducible randomness, incompleteness, and inconsistencies permeate elementary number theory, arithmetic, and Newtonian mechanics (Chaitin, 2003); and
– that longstanding calls for clearly distinguishing levels of complexity (Rousseau, 1985; Star & Ruhleder, 1996, p. 118) typically go unheeded.

We explicitly call for explorations of ways to separate levels of complexity in the measurement context and applaud recent efforts in this vein that expand on Galison's notion of the trading zone and Star's theory of the boundary object (Confrey et al., 2021; Fisher & Wilson, 2015; Lehrer & Jones, 2014). These efforts expand on Galison's (1997) documentation of the complex nonlinearities he found exhibited across the discontinuously interrelated microphysics communities of experimentalists, instrument makers, and theoreticians. Independent support for Galison's sense of the paradoxical positive functionality produced when convergent agreement is complemented by some kinds of divergent disagreement is provided by Woolley and Fuchs' (2011) study of collective intelligence in the organization of science.

Additional support is evident in Ostrom's theory of institutional organization, where a nested hierarchy of concrete operational rules, abstract collective-choice rules, and formal constitutional rules are distinguished: "Constitutional-choice rules affect operational activities and results through their effects in determining who is eligible and determining the specific rules to be used in crafting the set of collective-choice rules that in turn affect the set of operational rules" (Ostrom, 2015, p. 52; Kiser & Ostrom, 1982). We expect that our research results will make substantive contributions to furthering Ostrom's program of participatory social ecologies, in the manner described by Fisher and Stenner (2018).

We are especially focused on applications where significant portions of the population exhibit different sensitivities in discriminating differences (Melin et al., 2022a). Feedback on these differences may comprise concretely actionable information useful to end-users and so ought to be systematically reported to them in common languages and formats throughout interconnected, quality assured metrological networks (Fisher, Oon, & Benson, 2021; Mallinson, 2024; Penuel et al., 2016, 2020).

This methodology differs from that employed in IRT in that measurements are not assumed to reduce population characteristics in a homogenizing, deductive way, necessitating either the forcing of round pegs into square holes, or uncontrollable variation in unit definitions. Instead, because the measurement model is not meant to be true, but must be useful (Rasch, 1960, pp. 37–38; Rasch, 1972/2011; Box, 1979), and in accord with the idea that measurement extends and feeds back into everyday language (Fisher, 2020, 2023), standards are seen as mediating inherently unrealistic formal axioms and locally idiosyncratic concrete circumstances. We aim to revitalize dialogue at the point of use as a means by which ambiguities are reconciled and shared points of reference are negotiated, as when a request to "open a window" has to be clarified by mentioning the stuffy room, or pointing at a computer screen.

Mediating standards operationalize the substantive value and enhanced defensibility obtained vis-a-vis individualized inferences when theoretical explanations and empirical estimates of person and item locations are predictable, repeatable, and reproducible. The "black box" of empirical analyses demonstrating separable parameters in single instances lacking defined constructs is insufficient to the task of scientific measurement. Substantive understanding must be demonstrated via theoretical explanations and predictions.

The integration of formal and concrete levels of complexity in abstract measurements is then further augmented by restricting inferences so that the information represented is associated with and derived from the organizational level the data describe. We agree here with the hypothesis offered by Hayman, Rayder, Stenner, and Madey (1978, p. 31) that, "the closer a set of data is to the organizational level for which it will be used (for decision-making), the more useful the data will be." Thus, treating counts of correct responses or summed ratings as measurements commits the ecological fallacy (Alker, 1969; Gnaldi et al., 2018) by mistaking numbers for quantities (Fisher, 2021). Reporting only interval measurements to end-users invested in the concrete application of the original data then also obscures the very information on responses most vital to their decision processes.

Information on variation in item discrimination is not ignored at the abstract level of the measurements, of course, since it correlates very highly with commonly employed model fit statistics (see figure 2 in Wright, 1995) and can be reported for every item and every category transition threshold using software like Winsteps (Linacre, 2023). For examples of end-user reports illustrating statistical and graphical representations of individual anomalies, see figure 8.8.2 and table 8.8.1 in Wright and Stone (1979, pp. 207–208). Reporting concrete exceptions that prove the rule to end-users could productively complement the reporting of abstract SI unit measurements and formal explanatory CSE predictions.

In line with others, we have argued for the significance of proper designs of measurement. Although today's society already has huge access to data, it is expected to continue to increase. This opens up the need for even more explorative approaches to understanding new phenomena and networks. At the same time, even when taking a more explorative approach, neither latent traits themselves nor measurements of them happen purely by accident. It cannot only be an empirical matter with a conception that views validity as something to be discovered afterward. However, we can learn from empirical studies how these new ideas can be expressed in relation to existing ones (Fisher, Melin, & Möller, 2021). To quote Andrich and Marais (2019): *when the data do not accord with the model, then the model can still be very useful in understanding the data. It helps to diagnose where the data are different from what was expected from the model. Usually, there is an explanation for such effects.* Therefore, combining theory-driven designs with an open-minded explorative approach could help when seeking new units to extend the SI. For example, available data may be submitted to a measurement model-based analysis as a way of leveraging low-hanging fruit capable of indicating the possibility of defining a potential new item hierarchy, one that might consequently be rearticulated as a CSE.

CSEs might also be used as an explorative tool for advancing the understanding of a latent trait itself and, consequently, the latent trait as measured. A CSE provides a more specific, causal, and rigorously mathematical conceptualization of latent traits than any other construct theory (Melin, Cano, & Pendrill, 2021). Our previous work has suggested three key parts in selecting explanatory variables to be included in a CSE (Pendrill, 2019; Melin & Pendrill, 2023), which also can be seen as an exploration toward a better understanding of the latent trait itself. First, it must be a conceptual, practical, or clinical judgment to define appropriate explanatory variables to be tested. Secondly, statistical guidance is needed to find the most significant explanatory variables to be included in the CSE. For guidance, a univariate correlation study between the latent trait of interest and each explanatory variable is complemented, in the PCR, by a multivariate correlation matrix formulated to evaluate the degree of correlation and the intercorrelations between the explanatory variables. Thirdly, when developed in a PCR, the performance of the CSE itself, as well as the amount of contribution from each explanatory variable, to the latent traits as measured against $\boldsymbol{X}'$ in terms of the principal components is evaluated by (i) the strength of correlation between the prediction (i.e., $\theta_{i,z}$ or $\delta_{j,z}$) and the latent traits as measured (i.e., $\theta_{i,m}$ or $\delta_{j,m}$) and (ii) the dispersion of the $\beta$-coefficients of the CSE (Melin & Pendrill, 2023). Thus, by adding or removing variables, one can use the CSE as an exploratory tool to advance the understanding of the latent trait operationalized in the construct theory. Again, a good qualitative understanding of the latent trait to be measured and a good fit to the model are preconditions to avoid misleading CSEs and interpretations of what is causing variation in the latent trait itself attributed to a person $\theta_i, g(\chi_i, \varepsilon_i, \beta)$ and a latent trait attributed to a task $\delta_j, f\left(\upsilon_j, \epsilon_j, \gamma\right)$ (Figures 5.2 and 5.8).

Building on the work by Adroher and Tennant (2019), who used clinical judgments to explain variation in task difficulty in activities of daily living, and Fisher (2012), who rated variations in physical functioning items, CSEs formulated with *qualitative* explanatory variables may also be possible. In two recent studies, we have explored this for balance measurements (Melin et al., 2023c) and upper limb measurements (Wangdell et al., 2023). In both works, explanatory variables that linearly increase or decrease along the continuum of either balance task difficulty or upper limb task difficulty were identified. Importantly, one must seek the demands required for the tasks themselves, not for a specific person/group of persons performing them. In a second stage, healthcare professionals were invited to score each of the identified explanatory variables for each item in the Berg Balance Scale or Tetraplegia Upper Limb Questionnaire. Note that these were not the same healthcare professionals for both cases; they were recruited for each study with specific domain expertise, in a manner related to that described by Bunderson et al. (2009). Subsequent analysis of the scored explanatory variables provided estimates of linear interval measures for each variable that subsequently could be used in the CSE. While both studies have shown methodological and conceptual possibilities, several concerns to be considered in further work have been highlighted.

In both studies, we have also discussed the role of entropy as in our earlier NeuroMET studies. At the same time, one must remember that there must be general demands on the body to perform different tasks, which differs from explaining an individual person's ability (Melin et al., 2023; Wangdell et al., 2023). Secondly, one may use a group of people to define the explanatory variables; it is likely that a group of people who can all perform all tasks equally well will have a very low variation in an entropy measure, and the average entropy is expected to increase linearly with the difficulty of the tasks. We hope those works open for further discussion and investigations to advance measurement quality assurance by including CSEs as a means of validity for understanding the latent trait intended to be measured.

Finally, cross-disciplinary efforts when measuring latent traits are warranted. Potential key roles have been presented in Section 5.3.3, but we highlight the significance of developing structures and forums for such cross-disciplinary efforts here. For example, the EU-funded *Measuring the impossible* (Pendrill et al., 2010) could be seen as a forerunner where different disciplines met somewhere between psychology and engineering to advance measurement methods for the human and social sciences. Thus, a better understanding of validity when extending models, measurement, and metrology of the SI cannot be an isolated activity only within or only outside the metrological community.

### 5.4.2 Validity claims today and tomorrow

With exponential growth in society's need to make well-informed decisions based on latent traits, and at the same time, from a strict metrological perspective with undeveloped models and methods, the understanding of the latent trait themselves and practical tools and advanced methodologies to measure the latent traits must be developed simultaneously. Today, weaker validity claims than tomorrow must be allowed, and for some latent traits of interest, weaker validity claims than others must be allowed. Nevertheless, with weaker validity claims, this needs to be communicated transparently, and, in turn, responsibility must be taken for the consequences of decisions being made based on those claims.

Inspired by physical metrology and centuries of continuously improving the measurements, we must be dedicated to advancing methods for meeting society's needs for fit-for-purpose and high-quality measurements of latent traits. This includes not stopping with "our job" when finding a set of items fitting an appropriate measurement model. Rather, one must continue to test in new and diverse samples and cross-country studies, evaluate possibilities when adding items to improve targeting and reliability without jeopardizing validity, and so on. On a global level, this also includes establishing and coordinating metrological references to support comparable measurement values of latent traits over time or between different areas. Thus, by continuously improving our methods, tomorrow's validity claims will be stronger for latent traits.

## 5.5 Conclusion

When positioning models, measurement, and metrology to extend the SI, the concept of validity is essential. It is hoped that this review and discussion about validity and its related aspects in the human and social sciences will contribute to including a more nuanced view of validity. However, we have not provided and have no intention of providing a panacea or a one-size-fits-all route for better use of validity terminology and validity. Rather we have proposed different routes, originating from the three important claims of validity to distinguish, and we hope it will stimulate a fresh look at what might be possible.

Overall, claims about the validity of the latent traits themselves, the validity when measuring latent traits, and the validity in decisions about latent traits and methods for validation should not be mixed. Careful and responsible actions must be taken when making claims about measurement-related validity and decision-related validity before validity in the latent trait itself is ensured. However, it is important to remain open for an iterative, explorative, and cross-disciplinary effort to advance both the validity of the latent traits themselves and the measured latent traits and develop methods for validation of latent traits themselves and the measured latent traits.

# References

Adroher, N. D., & Tennant, A. (2019). Supporting construct validity of the evaluation of daily activity questionnaire using linear logistic test models. *Quality of Life Research*, *28*(6), 1627–1639, https://doi.org/10.1007/s11136-019-02146-4

Alker, H. R. (1969). A typology of ecological fallacies. In M. Dogan & S. Rokkan (Eds.). *Quantitative ecological analysis in the social sciences* (pp. 69–86). MIT Press.

Andersen, E. B. (1999). Sufficient statistics in educational measurement. In G. N. Masters & J. P. Keeves (Eds.). *Advances in measurement in educational research and assessment* (pp. 122–125). Pergamon.

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*, 449–460.

Andrich, D. (1988). *sage university paper series on quantitative applications in the social sciences.* Vol. series no. 07–068: Rasch models for measurement. Sage Publications.

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika*, *75*(2), 292–308.

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer Texts in Education, Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-7496-8

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions*, *22*(1), 1145–1146.

Barbic, S. P., Cano, S. J., & Mathias, S. (2018). The problem of patient-centred outcome measurement in psychiatry: Why metrology hasn't mattered and why it should. *Journal of Physics: Conference Series*, *1044*, 012069. https://doi.org/10.1088/1742-6596/1044/1/012069

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of pair comparisons. *Biometrika*, 63, 324–345.

De Battisti, F., Nicolini, G., & Salini, S. (2010). The Rasch model in customer satisfaction survey data. *Quality Technology & Quantitative Management. Taylor & Francis*, *7*(1), 15–34, https://doi.org/10.1080/16843703.2010.11673216

Bentley, J. P. (2005). *Principles of measurement systems*. 4th ed., Harlow, England ; New York: Pearson Prentice Hall.

Bohm, D. (1952). A suggested interpretation of the quantum theory in terms of "hidden" variables. I. *Physical Review*, *85*(2), 166–179.

Bohm, D. (2005). *Wholeness and the implicate order*. Routledge.

Bohm, D., & Hiley, B. J. (1984). Measurement understood through the quantum potential approach. *Foundations of Physics*, *14*(3), 255–274.

Bohm, D., & Hiley, B. J. (1989). Non-locality and locality in the stochastic interpretation of quantum mechanics. *Physics Reports*, *172*(3), 93–122.

Bohm, D., & Hiley, B. J. (2006). *The undivided universe: An ontological interpretation of quantum theory*. Routledge.

Bohm, D., Hiley, B. J., & Kaloyerou, P. N. (1987). An ontological basis for the quantum theory. *Physics Reports*, *144*(6), 321–375.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511490026

Denny, B., 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

Denny, B., 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

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.). *Robustness in statistics* (pp. 201–235). Academic Press, Inc.

Bunderson, C. V., & Newby, V. A. (2009). The relationships among design experiments, invariant measurement scales, and domain theories. *Journal of Applied Measurement*, *10*(2), 117–137.

Chaitin, G. J. (1994). Randomness and complexity in pure mathematics. *International Journal of Bifurcation and Chaos*, *4*(1), 3–15, http://www.worldscientific.com/doi/pdf/10.1142/S0218127494000022

Chaitin, G. J. (2003). The limits of mathematics. Springer-Verlag.

Clifton, J. D. W. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods*, *25*(3), 259–270, https://doi.org/10.1037/met0000236

Confrey, J., Shah, M., & Toutkoushian, E. (2021). Validation of a learning trajectory-based diagnostic mathematics assessment system as a trading zone. *Frontiers in Education: Assessment, Testing and Applied Measurement*, *6*(654353), doi:10.3389/feduc.2021.654353

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

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Statistics for Social and Behavioral Sciences. Springer-Verlag.

Dybkaer, R. (2010). ISO terminological analysis of the VIM3 concepts 'quantity' and 'kind-of-quantity. *Metrologia*, *47*(3), 127, https://doi.org/10.1088/0026-1394/47/3/003

Embretson, S. E. (1996). Item Response Theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201–212.

Embretson, S. E. (2010). *Measuring psychological constructs: Advances in model-based approaches*. American Psychological Association.

Engelhard, G., Jr (2012). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge Academic.

Finkelstein, L. (1975). Fundamental concepts of measurement: Definition and scales. *Measurement and Control*, *8*(3), 105–111. SAGE Publications Ltd, https://doi.org/10.1177/002029407500800305

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement*, *34*(1), 39–48, https://doi.org/10.1016/S0263-2241(03)00018-6

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

Fisher, W. P., Jr (1988). *Truth, method, and measurement: The hermeneutic of instrumentation and the Rasch model*. Dissertation, Dissertation Abstracts International (University of Chicago, Dept. of Education, Division of the Social Sciences). 49, 0778A.

Fisher, W. P., Jr. (1994). The Rasch debate: Validity and revolution in educational measurement. In, 36–72.

Fisher, W. P., Jr (2020). Contextualizing sustainable development metric standards: Imagining new entrepreneurial possibilities. *Sustainability*, *12*(9661), 1–22, https://doi.org/10.3390/su12229661

Fisher, W. P., Jr (2021). Bateson and Wright on number and quantity: How to not separate thinking from its relational context. *Symmetry*, *13*(1415), https://doi.org/10.3390/sym13081415

Fisher, W. P., Jr (2023). Measurement systems, brilliant results, and brilliant processes in healthcare: Untapped potentials of person-centered outcome metrology for cultivating trust. In W. P. Fisher Jr. & S. Cano (Eds.). *Person-centered outcome metrology* (pp. 357–396). Springer, https://link.springer.com/book/10.1007/978-3-031-07465-3

Fisher, W. P., Jr (2012). A predictive theory for the calibration of physical functioning patient survey items. *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.2084490

Fisher, W. P., Jr, Melin, J., & Möller, C. (2021). *Metrology for climate-neutral cities*. http://urn.kb.se/resolve?urn=urn:nbn:se:ri:diva-57281. (12 September, 2022).

Fisher, W. P., Jr., Oon, E. P.-T., & Benson, S. (2021). Rethinking the role of educational assessment in classroom communities: How can design thinking address the problems of coherence and complexity? *Educational Design Research*, *5*(1), 1–33.

Fisher, W. P., Jr, & Jackson Stenner, A. (2011). Integrating qualitative and quantitative research approaches via the phenomenological method. *International Journal of Multiple Research Approaches*, *5*(1), 89–103, https://doi.org/10.5172/mra.2011.5.1.89

Fisher, W. P., Jr, & Jackson Stenner, A. (2018). Ecologizing vs modernizing in measurement and metrology. *Journal of Physics Conference Series*, *1044*(012025), http://iopscience.iop.org/article/10.1088/1742-6596/1044/1/012025

Fisher, W. P., Jr, & Wilson, M. (2015). Building a productive trading zone in educational assessment research and practice. *Pensamiento Educativo: Revista de Investigacion Educacional Latinoamericana*, *52*(2), 55–78, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2688260

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. In *Advances in methods and practices in psychological science*. SAGE Publications Inc. https://doi.org/10.1177/2515245920952393

Galison, P. (1997). *Image and logic: A material culture of microphysics*. University of Chicago Press.

Geisinger, K. F. (1992). The metamorphosis to test validation. *Educational Psychologist*, *27*(2), 197–222, https://doi.org/10.1207/s15326985ep2702_5

Gnaldi, M., Tomaselli, V., & Forcina, A. (2018). Ecological fallacy and covariates: New insights based on multilevel modelling of individual data. *International Statistical Review*, *86*(1), 119–135.

Esfeld, M., Lazarovici, D., Hubert, M., & Dürr, D. (2014). The ontology of Bohmian mechanics. *The British Journal for the Philosophy of Science*, *65*(4), 773–796, http://www.jstor.org/stable/24562842

Goldstein, S. (1998a). Quantum theory without observers-Part one. *Physics Today*, *51*(3), 42–47.

Goldstein, S. (1998b). Quantum theory without observers-Part two. *Physics Today*, *51*(4), 38–42, https://doi.org/10.1063/1.882241

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing: *Measurement and Evaluation in Counseling and Development*. *36*(3), 181–191. *Routledge*, https://doi.org/10.1080/07481756.2003.11909741

Hambleton, R. K., Swaminathan, H., & Rogers, L. (1991). *Fundamentals of item response theory*. Sage Publications.

Harvey, P., Jensen, C. B., & Morita, A. (2017). *Infrastructures and social complexity: A companion*. Taylor & Francis.

Hayman, J., Rayder, N., Stenner, A. J., & Madey, D. L. (1979). On aggregation, generalization, and utility in educational evaluation. *Educational Evaluation and Policy Analysis*, *1*(4), 31–39.

JCGM [20]0. (2012). *International vocabulary of metrology – Basic and general concepts and associated terms (VIM)*. BIPM.

Johansson, M., Preuter, M., Karlsson, S., Möllerberg, M.-L., Svensson, H., & Melin, J. (2023). *Valid and Reliable? Basic and Expanded Recommendations for Psychometric Reporting and Quality Assessment*. OSF Preprints, https://doi.org/10.31219/osf.io/3htzc

Joint Committee on the Standards for Educational and Psychological Testing. (2014). *Standards for Educational and Psychological Testing*. Washington, D.C: American Educational Research Association.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 198–211, https://doi.org/10.1080/0969594X.2015.1060192

Karlsson, S., Melin, J., Svensson, H., & Wisén, J. (2023). *A metrological approach to social sustainability metrics in municipalities*. OSF Preprints, https://doi.org/10.31219/osf.io/sdzwn

Kelley, T. L. (1927). *Interpretation of educational measurements.* (Interpretation of Educational Measurements). Oxford, England: World Book Co.

Kiser, L. L., & Ostrom, E. (1982). The three worlds of action: A metatheoretical synthesis of institutional approaches. In E. Ostrom (Ed.). *Strategies of political inquiry* (pp. 179–222). Sage.

Koopmans, T. C. (1947). Measurement without theory. *The Review of Economics and Statistics*, *29*(3), 161, https://doi.org/10.2307/1928627

Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change*. Revised. edition, Chicago, Ill: University of Chicago Press.

Lehrer, R. (2013). A learning progression emerges in a trading zone of professional community and identity. *WISDOMe Monographs*, *3*, 173–186.

Lehrer, R., & Jones, S. (2014, April 2). Construct maps as boundary objects in the trading zone. In W. P. Fisher Jr. (Chair). *Session 3-A: Rating scales and partial credit, theory and applied*. Philadelphia, PA: International Objective Measurement Workshop.

Linacre, J. M. (1995). Paired comparisons with ties: Bradley-Terry and Rasch. *Rasch Measurement Transactions*, *9*(2), 425, http://www.rasch.org/rmt/rmt92d.htm

Linacre, J. M. (2000a). Was the Rasch model almost the Peirce model? *Rasch Measurement Transactions*, *14*(3), 756–757, http://www.rasch.org/rmt/rmt143b.htm

Linacre, J. M. (2000b). Almost the Zermelo model? *Rasch Measurement Transactions*, *14*(2), 754, http://www.rasch.org/rmt/rmt142k.htm

Linacre, J. M. (2023). *A user's guide to WINSTEPS Rasch-Model computer program, v. 5.6.2*. Winsteps.com, https://www.winsteps.com/manuals.htm

Lovejoy, D. (1999). Objectivity, causality, and ideology in modern physics. *Science & Society*, *63*(4), 433–468.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review, 66*(2), 81–95.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new kind of fundamental measurement. *Journal of Mathematical Psychology*, 1(1), 1–27.

Mallinson, T. (2024). Extending the justice-oriented, anti-racist framework for validity testing to the application of measurement theory in re(developing) rehabilitation assessments. In W. P. Fisher Jr. & L. Pendrill (Eds.). *Models, measurement, and metrology extending the SI*. in press, De Gruyter.

Mari, L., Wilson, M., & Maul, A. (2022). *Measurement across the sciences: Developing a shared concept system for measurement*. 1st ed. 2021 edition, Springer.

Matarese, V. (2023). *Epistemic studies. Vol. 51: The metaphysics of Bohmian mechanics: A comprehensive guide to the different interpretations of Bohmian ontology*. M. Esfeld, S. Hartmann, & A. Newen (Eds.). De Gruyter.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69, https://doi.org/10.1080/15366367.2017.1348108

Maul, A., Torres Irribarra, D., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement*, *79*, 311–320. https://doi.org/10.1016/j.measurement.2015.11.001

McAllister, S. (2008). Introduction to the use of Rasch analysis to assess patient performance. *International Journal of Therapy & Rehabilitation*, *15*(11), 482–490. Mark Allen Holdings Limited, https://doi.org/10.12968/ijtr.2008.15.11.31544

McKenna, S. P., Heaney, A., & Wilburn, J. (2019). Measurement of patient-reported outcomes. 2: Are current measures failing us? *Journal of Medical Economics*, *22*(6), 523–530, https://doi.org/10.1080/13696998.2018.1560304

McKenna, S. P., Heaney, A., Wilburn, J., & Jackson Stenner, A. (2019). Measurement of patient-reported outcomes. 1: The search for the Holy Grail. *Journal of Medical Economics*, *22*(6), 516–522. Taylor & Francis, https://doi.org/10.1080/13696998.2018.1560303

Melin, J. (2021). Neurogenerative disease metrology and innovation: The European Metrology Programme for Innovation & Research (EMPIR) and the NeuroMET projects. Conference presentation presented at the Pacific Rim Objective Measurement Symposium 2021. https://proms.promsociety.org/2021/.

Melin, J., Cano, S., Flöel, A., Göschel, L., & Pendrill, L. (2022a). The role of entropy in construct specification equations (CSE) to Improve the validity of memory tests: Extension to word lists. *Entropy*, *24*(7), 934. Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/e24070934

Melin, J., Cano, S. J., Flöel, A., Göschel, L., & Pendrill, L. R. (2022b). Metrological advancements in cognitive measurement: A worked example with the NeuroMET memory metric providing more reliability and efficiency. *Measurement: Sensors*, 100658. https://doi.org/10.1016/j.measen.2022.100658

Melin, J., Cano, S. J., Gillman, A., Marquis, S., Flöel, A., Göschel, L., & Pendrill, L. R. (2023a). Traceability and comparability through crosswalks with the NeuroMET memory metric. *Scientific Reports*, *13*(1), 1–12. Nature Publishing Group, https://doi.org/10.1038/s41598-023-32208-0

Melin, J., Pendrill, L. R., & Cano, S. J. EMPIR NeuroMET 15HLT04 consortium. (2019) Towards patient-centred cognition metrics. *Journal of Physics: Conference Series*, 012029. https://doi.org/10.1088/1742-6596/1379/1/012029

Melin, J., Cano, S., & Pendrill, L. (2021). The role of entropy in construct specification equations (CSE) to improve the validity of memory tests. *Entropy*, *23*(2), 212. Multidisciplinary Digital Publishing Institute https://doi.org/10.3390/e23020212

Melin, J., Göschel, L., Hagell, P., Westergren, A., Flöel, A., & Pendrill, L. (2023b). Forward and backward recalling sequences in spatial and verbal memory tasks: What do we measure? *Entropy*, *25*(5), 813. Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/e25050813

Melin, J., Fridberg, H., Ekvall Hansson, E., Smedberg, D., & Pendrill, L. (2023c). Exploring a new application of construct specification equations (CSEs) and entropy: A pilot study with balance measurements. *Entropy*,

Melin, J., & Pendrill, L. (2022a). Humans as measurement instruments and Construct specification equations (CSE) in measurement systems. *BEAR Seminar*, https://files.bearcenter.org/video/Melin Pendrill_HumansEquationsMeasurementSystems_20221115.mp4 ((4 May, 2023)).

Melin, J., & Pendrill, L. (2022b). A novel metrological approach to a more consistent way of defining and analyzing memory task difficulty in word learning list tests with repeated trials. In *Proceedings of the RaPID Workshop – Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments – Within the 13th Language Resources and Evaluation Conference* (pp. 17–21). Marseille, France: European Language Resources Association, https://aclanthology.org/2022.rapid-1.3 13 January, 2023.

Melin, J., & Pendrill, L. R. (2023). The role of construct specification equations and entropy in the measurement of memory. In F. William P. Jr. & S. J. Cano (Eds.). *Person-centered outcome metrology: principles and applications for high stakes decision making (*Springer series in measurement science and technology) (pp. 269–309). Cham: Springer International Publishing, https://doi.org/10.1007/978-3-031-07465-3_10

Messick, S. (1989a). Validity. In *Educational measurement*.3rd ed. (The American council on education/ Macmillan series on higher education) (pp. 13–103). American Council on Education.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. American Educational Research Association, https://doi.org/10.3102/0013189X018002005

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23. American Educational Research Association, https://doi.org/10.3102/0013189X023002013

Messick, S. (1996). Validity and Washback in Language Testing. *ETS Research Report Series*, *1996*(1), i–18, https://doi.org/10.1002/j.2333-8504.1996.tb01695.x

Michell, J. (2021). “The art of imposing measurement upon the mind”: Sir Francis Galton and the genesis of the psychometric paradigm. *Theory & Psychology*, *26*.

Morel, T., & Cano, S. J. (2017). Measuring what matters to rare disease patients – Reflections on the work by the IRDiRC taskforce on patient-centered outcome measures. *Orphanet Journal of Rare Diseases*, *12*(1), 171, https://doi.org/10.1186/s13023-017-0718-x

Morel, T., Cleanthous, S., Andrejack, J., Barker, R. A., Blavat, G., Brooks, W., Burns, P., et al. (2022). Patient experience in early-stage Parkinson’s Disease: Using a mixed methods analysis to identify which concepts are cardinal for clinical trial outcome assessment. *Neurology and Therapy*, *11*(3), 1319–1340. https://doi.org/10.1007/s40120-022-00375-3

Lim, M., Sok, S. R., & Brown, T. (2009). Using Rasch analysis to establish the construct validity of rehabilitation assessment tools. *International Journal of Therapy and Rehabilitation*, *16*(5), 251–260, https://doi.org/10.12968/ijtr.2009.16.5.42102

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. V. (2009). The Rasch model and additive conjoint measurement. *Journal of Applied Measurement*, *10*(4), 348–354.

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational & psychological assessment*. Vol. 55, London: City Road. https://doi.org/10.4135/9781446288856.

Ostrom, E. (2015). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press. Original work published 1990.

Peirce, C. S. (1878). Illustration of the logic of science. Fourth paper: The probability of induction. *Popular Science Monthly*, *12*, 705–718. (Rpt N. Houser & C. Kloesel (Eds.). 1992, *The essential Peirce: Selected philosophical writings, vol. I*. 1867–1893, (pp. 155–169), Indiana University Press.).

Pendrill, L. (2014). Man as a measurement instrument. *NCSLI Measure*, *9*(4), 24–35, https://doi.org/10.1080/19315775.2014.11721702

Pendrill, L. (2019). *Quality assured measurement: Unification across social and physical sciences*. (Springer Series in Measurement Science and Technology). Springer International Publishing, https://doi.org/10.1007/978-3-030-28695-8

Pendrill, L. (2018). Assuring measurement quality in person-centred healthcare. *Measurement Science and Technology*, *29*(3), 034003, https://doi.org/10.1088/1361-6501/aa9cd2

Pendrill, L. (2021). Quantities and units in quality assured measurement. Presented at the PACIFIC RIM OBJECTIVE MEASUREMENT SYMPOSIUM 2021. https://proms.promsociety.org/2021/.

Pendrill, L. (2024). Quantities and units: Order amongst complexity. In W. P. Fisher, Jr. & L. R. Pendrill (Eds.), *Models, measurement, and metrology extending the SI*, (pp. 35–100). De Gruyter.

Pendrill, L. R., Emardson, R., Berglund, B., Gröning, M., Höglund, A., Cancedda, A., Quinti, G., et al. (2010). Measurement with persons: a European network. *NCSLI Measure*, *5*(2), 42–54. Taylor & Francis, https://doi.org/10.1080/19315775.2010.11721515

Penuel, W. R., Clark, T. L., & Bevan, B. (2016). Infrastructures to support equitable STEM learning across settings. *After School Matters*, *24*, 12–20.

Penuel, W. R., Riedy, R., Barber, M. S., Peurach, D. J., LeBouef, W. A., & Clark, T. (2020). Principles of collaborative education research with stakeholders: Toward requirements for a new research and development infrastructure. *Review of Educational Research*, *90*(5), 627–674.

Prigogine, I. (1971). Unity of physical laws and levels of description. In I. Prigogine & M. Grene (Eds.). *Interpretations of life and mind: Essays around the problem of reduction* (pp. 1–13). Humanities Press.

Prigogine, I. (1976). Order through fluctuation: Self-organization and social system. In E. Jantsch & C. Waddington (Eds.). *Consciousness and evolution: Human systems in transition* (pp. 93–130). Addison Wesley.

Prigogine, I. (1978). Time, structure and fluctuations [Nobel lecture]. *Science*, *201*, 777–785.

Prigogine, I., & Stengers, I. (2018). *Order out of chaos: Man's new dialogue with nature*. Verso.

Quaglia, M., Cano, S., Fillmer, A., Flöel, A., Giangrande, C., Göschel, L., Lehmann, S., Melin, J., & Teunissen, C. E. (2021). The NeuroMET project: Metrology and innovation for early diagnosis and accurate stratification of patients with neurodegenerative diseases. *Alzheimer's & Dementia*, *17*(S5), e053655. John Wiley & Sons, Ltd, https://doi.org/10.1002/alz.053655

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.

Rasch, G. (1966). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.). *Readings in mathematical social science* (pp. 89–108). Science Research Associates, https://www.rasch.org/memo19662.pdf

Rasch, G. (1973/2011). All statistical models are wrong! Comments on a paper presented by Per Martin-Löf, at the Conference on Foundational Questions in Statistical Inference, Aarhus, Denmark. *Rasch Measurement Transactions*, *24*(4), 1309. May 7–12, 1973, http://www.rasch.org/rmt/rmt244.pdf

Rousseau, D. M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. *Research in Organizational Behavior*, *7*(1), 1–37.

San Martin, E., Gonzalez, J., & Tuerlinckx, F. (2009). Identified parameters, parameters of interest, and their relationships. *Measurement: Interdisciplinary Research and Perspectives*, *7*(2), 97–105.

San Martin, E., Gonzalez, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3 PL model. *Psychometrika*, *80*(2), 450–467.

San Martin, E., & Rolin, J. M. (2013). Identification of parametric Rasch-type models. *Journal of Statistical Planning and Inference*, *143*(1), 116–130.

Slaney, K. (2017). *Validating psychological constructs*. London: Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-38523-9

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1), 111–134.

Stenner, A. J. (2014). Validity revisited. *Presented at the IOMW – Philadelphia April*, *1*, 2014.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement*, *7*(3), 307–322.

Stenner, A. J., Fisher, W. P., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology*, *4*, https://doi.org/10.3389/fpsyg.2013.00536

Stenner, A. J., & Smith, M. (1982). Testing construct theories. *Perceptual and Motor Skills*, *55*(2), 415–426, https://doi.org/10.2466/pms.1982.55.2.415

Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, *20*(4), 305–316.

Tal, E. (2014). Making time: A study in the epistemology of measurement. *The British Journal for the Philosophy of Science*, *67*(1), 297–335, https://doi.org/10.1093/bjps/axu037

Tesio, L. (2003). Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *Journal of Rehabilitation Medicine*, 2003, 11.

Tesio, L., Caronni, A., Kumbhare, D., & Scarano, S. (2023). Interpreting results from Rasch analysis 1. The "most likely" measures coming from the model. *Disability and Rehabilitation*, 1–13. https://doi.org/10.1080/09638288.2023.2169771

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, XXXIII, 529–544. (Rpt L. L. Thurstone. 1959, *The measurement of values.* (pp. 215–233). University of Chicago Press, Midway Reprint Series).

Van Iddekinge, Chad, H., Lievens, F., & Sackett, P. R. (2023). Personnel selection: A review of ways to maximize validity, diversity, and the applicant experience. *Personnel Psychology* n/a(n/a). https://doi.org/10.1111/peps.12578.

Johanna, W., Pendrill, L., Dunn, J., Hill, B., & Melin, J. (2023). Construct specification equations to improve validity in upper limb measurements. *Frontiers in Rehabilitation Sciences*. https://assets.researchsquare.com/files/rs-4128671/v1_covered_4172cb71-d2ec-41c1-8f72-f5ccc01a2afa.pdf

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, N.J: Lawrence Erlbaum Associates.

Wolf, M. G., Ihm, E., Maul, A., & Taves, A. (2019). Survey item validation. In M. Stausberg & S. Engler (Eds.). *Handbook of Research Methods in the Study of Religion*. Vol. 10. 2nd ed., Routledge.

Woolley, A. W., & Fuchs, E. (2011). Collective intelligence in the organization of science. *Organization Science*, *22*(5), 1359–1367.

Wright, B. D. (1995). 3 PL IRT or Rasch? *Rasch Measurement Transactions*, *9*(1), 408, http://www.rasch.org/rmt/rmt91b.htm

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33–45, 52, https://doi.org/10.1111/j.1745-3992.1997.tb00606.x

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

Zermelo, E. (1929). The calculation of tournament results as a maximum-likelihood problem [German]. *Mathematische Zeitschrift*, *29*, 436–460.