GESIS Leibniz Institute for the Social Sciences

Validity in Survey Research – From Research Design to Measurement Instruments

Lydia Repke, Lukas Birkenmaier, & Clemens M. Lechner

July 2024, Version 1.0

Abstract

The ability to draw valid conclusions from data is crucial for any empirical research. Thus, validity is one of the leading quality criteria in the social and behavioral sciences. However, the term validity is used very differently across disciplines and time, creating terminological confusion that can render the concept elusive. This survey guideline provides a compact overview of different meanings associated with the term validity in the social and behavioral sciences. To acknowledge the term's full breadth, we first distinguish between (a) validity pertaining to the research design and (b) validity pertaining to measurement instruments. We show that validity is fundamentally about whether the research design and measurement instruments used for a study are true to what they are theoretically supposed to represent or capture. Subsequently, we focus on providing practical guidance on assessing measurement validity, that is, a measurement instrument's ability to measure what it purports to measure. In particular, we discuss the types of evidence supporting measurement validity and the methods researchers can use to provide such evidence for survey research. Our aim is to equip researchers with a conceptual understanding of measurement validity and a toolkit for assessing the validity of measurement instruments. We emphasize that validity is not a fixed property of a measurement instrument. Instead, researchers should view validity as a dynamic process of validation. This ongoing practice involves supporting and justifying conclusions drawn from survey data through a combination of theoretical reasoning and empirical evidence.

Citation

Repke, Lydia; Birkenmaier, Lukas; Lechner, Clemens M. (2024). Validity in Survey Research - From Research Design to Measurement Instruments. Mannheim, GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines).

DOI: 10.15465/gesis-sg_en_048

1. Introduction

In the social and behavioral sciences, researchers often conduct surveys to draw conclusions about populations and answer research questions about social phenomena. They thereby rely on measurement instruments (e.g., items, scales, tests) designed to assess manifest (observable) and latent (unobservable) characteristics of individuals, groups, organizations, or other entities. The better their research designs and measurement instruments are, the closer researchers' conclusions are to reality. It is not always clear, however, how good surveys are at measuring what they intend to measure. Therefore, attempts have been made to define and assess the quality of study designs, measurement instruments, and conclusions drawn from both.

While three main criteria determine the quality of scientific research designs and the assessment of measurement instruments—i.e., objectivity, reliability, and validity—, validity is the quality criterion that is the most fundamental one.[1] In essence, validity is the extent to which a scientific study or a specific measure captures what it is supposed to measure. Or, in other words, are researchers able to draw the conclusions they intend to draw, or do researchers measure what they intend to measure? Although the basic idea behind the concept of validity is relatively straightforward, the literature on validity theory has experienced remarkable transitions and fragmentation across academic disciplines and time (Borsboom, Mellenbergh & Heerden, 2004; Hughes, 2018; Kane, 2001), leading to much confusion and debate about what validity means in practice (Newton & Baird, 2016). In fact, the term validity encompasses different meanings and, as such, is used rather heterogeneously in the scientific community.

Historically, psychology, particularly psychometrics, has been a major contributor to the study of validity and its assessment. The field has produced extensive and innovative work in this area (Adcock & Collier, 2001). As a result, it is not surprising that psychology, among all social and behavioral disciplines, holds the longest tradition of employing the term validity and likely possesses the most nuanced understanding of it. However, the meaning attached to the term validity varies considerably, not only outside the field of psychology but also within it. This confusion begins with the fact that validity matters in different contexts: In research designs, validity denotes whether a study captures what it is supposed to measure, whereas, in the context of measuring a specific construct, validity refers to whether a variable or a set of variables assesses what it is supposed to measure.

Furthermore, there are many different validity-related terms (e.g., construct validity, face validity, concurrent validity), and each research field has implicit norms regarding assumptions and methods, making cross-disciplinary communication difficult (Adcock & Collier, 2001; McDermott, 2011; Birkenmaier, Lechner & Wagner, 2023). This problem gets amplified when scholars do not specify what they mean by validity or use the term incorrectly. For example, some researchers consider construct validity synonymous with overall validity, while others define construct validity as the accuracy with which a construct is labeled. Conversely, a specific validity characteristic may be known by various terms. For example, factorial validity refers to the same meaning as structural validity, and discriminant validity has the same meaning as divergent validity. In their comprehensive review, Newton and Shaw (2014) compiled a list of 150 validity-related terms scholars have proposed over the decades.

[1] Still, it is crucial to generate and use data that meet the standards of all three quality criteria: Even a highly objective and reliable measurement instrument is not very useful if it is not valid. In contrast, a valid measurement instrument can be equally problematic if it lacks reliability and objectivity. For more information on reliability, see Danner (2016).

Although we cannot review all the different meanings of validity in detail, this survey guideline provides a structured overview—akin to a glossary—of the most important meanings of the term validity and ways to assess them, primarily focusing on the validity of measurement instruments. Our goal is to equip researchers with the conceptual understanding and methodological toolbox needed to address the question of validity—an essential task for any research endeavor. We first summarize the aspects of validity related to the research design (Chapter 2) and then detail those related to measurement (Chapter 3). Ultimately, we deepen this topic by providing hands-on guidance on how to exemplarily validate a multi-item scale for measuring the latent concept of political trust (Chapter 4) and conclude our elaboration with further remarks and recommendations (Chapter 5).

2. Validity Considerations in Research Designs

When discussing the validity of a research design, three validity terms are commonly used: internal, external, and statistical (or conclusion or statistical conclusion) validity. First, internal validity pertains to making accurate claims about the cause and effect of the variables under study (Campbell & Stanley, 2015; Cook & Campbell, 1979). Essentially, it addresses whether an independent variable X is the cause of a dependent variable Y, assuming a relationship between the two. Hence, internal validity refers to the ability to draw conclusions about causal relationships based on the given data. In experimental designs, internal validity tends to be higher compared to non-experimental research designs due to the greater ease of controlled manipulation of the cause. Nonetheless, various threats to internal validity may still arise in experimental settings, including selection effects (i.e., study participants are not randomly assigned to experimental groups), mortality (i.e., withdrawals, dropouts, attrition), and testing issues (i.e., changes in test scores result from repeated testing and not from the intervention itself, learning effect; Slack & Draugalis, 2001).

Second, external validity concerns the extent to which the results or presumed causal relationships can be generalized across populations, settings, and time (Cook & Campbell, 1979). The underlying question here is whether the obtained results—assuming a causal relationship between the variables X and Y—are applicable to other contexts. Though important for any research, this is particularly crucial for cross-cultural research (e.g., Ward & Kennedy, 1994). For example, can findings regarding the political trust of Dutch immigrants in Spain also be extended to German immigrants? Establishing external validity necessitates the replication of studies across diverse populations and settings, ideally employing various methods and measurements (Aronson, Ellsworth, Carlsmith & Gonzales, 1990; McDermott, 2011). In contrast to internal validity, external validity is generally higher in natural (i.e., non-experimental) settings.

A specific form of external validity worth noting is ecological validity, which refers to the extent to which study findings generalize to settings or populations typical in today's society (Aronson, Wilson & Brewer, 1998). This form of external validity becomes particularly relevant when considering how well experimental results can be generalized to situations outside the controlled laboratory environment (i.e., real-world settings). Imagine, for instance, a laboratory experiment examining new teaching and learning methods and their effect on students' learning outcomes. To assess the ecological validity of this experiment, researchers could replicate the study in an actual classroom environment with diverse students and teachers; as this is not a controlled setting, factors such as classroom dynamics, teacher-student interactions, and so on could influence the results. If the results in the classroom setting align with those of the laboratory, however, it increases the confidence in the ecological validity of the findings.

Scholars commonly assume that internal and external validity are inversely related (Jiménez-Buedo & Miller, 2010). This understanding suggests that as one increases, the other tends to decrease. For example, scholars may perceive field experiments as having high external but low internal validity, while viewing laboratory experiments as having high internal but low external validity. However, the relationship between the two is not necessarily a strict trade-off (Jiménez-Buedo & Miller, 2010). For instance, in experiments, threats to internal and external validity related to inferential problems can often be addressed by replicating the experiment with slight variations (Jiménez-Buedo & Miller, 2010). In this sense, repeating the same experiment in different settings with diverse groups and incentives can lead to both higher internal and external validity.

Nevertheless, there are some instances where maximizing internal validity may be more important (e.g., when researchers want to be confident that the effect found resulted from their manipulation), while in other cases, the emphasis should be on external validity (e.g., when researchers want to evaluate how generalizable an effect is; see McDermott, 2011). Interestingly, there are disciplinary differences in the importance assigned to internal and external validity. While psychologists tend to focus on internal validity, political scientists are generally more concerned with external validity (McDermott, 2011). This discrepancy can be attributed partly to differences in research foci and purposes, with psychologists more often using experimental designs to test theories and political scientists focusing on broader generalizations across populations (McDermott, 2011).

Lastly, the terms statistical, conclusion, and statistical conclusion validity all refer to whether reasonable conclusions can be drawn from data analysis (Cook & Campbell, 1979). The underlying question is whether a relationship exists between the variables X and Y. Assessing statistical conclusion validity typically requires expertise or extensive training to determine whether it is plausible to assume the covariation of X and Y given a specific alpha level and variances. Factors essential to determining this aspect of validity include statistical power, adherence to statistical assumptions, and reliability of the measurements (Drost, 2004).

3. Validity Considerations in Measurement

Beyond validity considerations in research designs, the term validity is important and widely used in the context of measurement (i.e., the operationalization of variables, for example, in the form of survey questions). In the social and behavioral sciences, researchers commonly collect data for two types of variables: manifest (observable) and latent (unobservable) variables. Manifest variables are directly observable and refer to factual or objective variables for which information could, in principle, be acquired from sources other than the respondent (e.g., a person's age could be determined from the birth certificate). In contrast, latent variables are not directly observable and represent theoretical concepts (e.g., identity, power, trust, and intelligence). To measure such theoretical concepts, researchers "translate" them into constructs, that is, into sets of interrelated variables or clusters of variables that covary and are measurable (for further elaboration on the differentiation between concept and construct, see Harkness, Edwards, Hansen, Miller, Villar, 2010; Markus, 2008; Podsakoff, MacKenzie, Podsakoff, 2016).

For example, the theoretical concept of 'political trust' can be measured with questions capturing trust in various entities such as the government, parliament, political parties, and local government (see Figure 1; for alternative models, see Schneider, 2017). Oftentimes, multiple indicators are necessary to measure a latent variable. This is because no single indicator covers a construct in its full definitional breadth, and each indicator is subject to measurement error, thus requiring multiple indicators to obtain sufficiently reliable measures (such as the example of political trust). However, in some cases, multiple indicators are not necessary; sometimes, single indicators suffice or may be the only feasible option (e.g., measuring demographic variables, importance, or evaluations). Even for single-indicator measures, it can be helpful to distinguish between the latent variable and its observed indicator to highlight the fact that each indicator represents an imperfect depiction of the target construct.

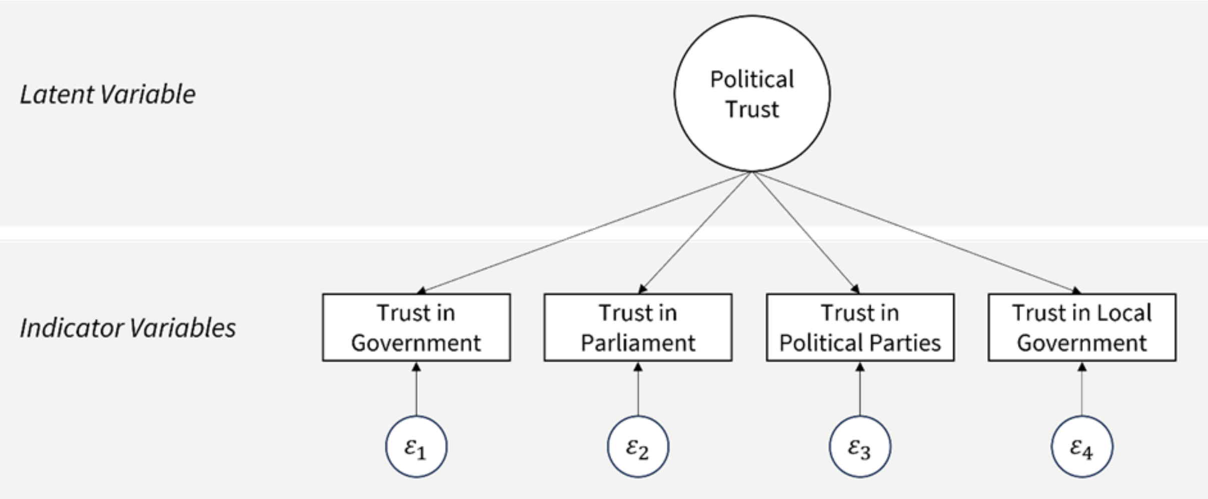

Figure 1: Example of the Latent Variable 'Political Trust' and Possible Indicator Variables

The ultimate goal of measuring either type of variable, be it observable or latent, is to draw meaningful conclusions about populations to answer questions about social phenomena. Construct validity is indispensable to demonstrate the trustworthiness of a measurement instrument credibly. It encompasses multiple kinds of validity-supporting evidence relevant to the interpretation or the meaning of respondents' answers to survey questions (i.e., measurement instruments; Messick, 1994). Validating a measurement instrument (i.e., the process of construct validation) entails accumulating multiple types of evidence to establish a sound scientific foundation for the validity of the construct's operationalization (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Today's understanding of construct validity as encompassing all kinds of validity-supporting evidence pertaining to the measurement instrument has not always prevailed. For several decades, a threefold typology dominated, wherein construct validity constituted only one of three subtypes of validity. Initially, in the early 20th century, validity was defined as a test property and described the extent to which a test (i.e., a standardized assessment of individuals' abilities, traits, behaviors, or attitudes) measures what it purports to measure (Kelley, 1927; Ruch, 1924). This early understanding of validity included, most prominently, criterion and content validity. While criterion validity relates to comparing the measurement scores against external criterion scores, content validity refers to selecting survey items based on theoretical considerations (Cureton, 1951; Hughes, 2018; Kane, 2001). Later, in the 1950s, Cronbach and Meehl (1955) introduced the idea of construct validity, which they understood—in contrast to today's view—as an additional but distinct type of validity that is involved whenever a test is to be interpreted as a measure of an attribute or a quality which is not operationally defined (i.e., a clear and specific procedure for measuring the construct is still lacking). In their view, construct validity emphasized the measurement process as a whole and thus required extended efforts toward interpreting theoretical constructs in specific research contexts.

This threefold typology of criterion, content, and construct validity was the dominant view for testing validity issues until some scholars challenged it (Adcock & Collier, 2001). They argued that dividing validity into subtypes had led to a simplification and hollowing out of the validity term, thereby narrowing and limiting validity questions to a checklist of mutually exclusive validity types (Goodwin, 2002). Thus, in the 1970s and 1980s, a unified approach to validity emerged (Newton, 2012). In particular, Messick (1994) coined a consensus definition of validity under the overarching concept of construct validity. In doing so, he called for referring to multiple types of validity-supporting evidence (such as content-related or criterion-related evidence) to demonstrate construct validity (Peetres & Harpe, 2020; Goodwin, 2002). This view emphasizes the importance of the research context when evaluating construct validity and might also include a cautious evaluation of the consequences of scientific interpretations and decisions based on test scores (Hughes, 2018).

Until today, this unified approach to validity has remained the dominant view. However, considerable differences still exist in the interpretation of construct validity between disciplines and research contexts. Some disciplines, for example, evaluate construct validity solely as a property of test scores or even interpretations of test scores (as, for example, in applied psychological testing), while others acknowledge that construct validity might also include the properties of a measurement instrument (Goodwin, 2002; Hughes, 2018). In this guideline, we endorse the first perspective. We view validity as an evaluative judgment on the interpretation of test scores, which rests on different kinds of validity-supporting evidence within a specific research context. Subsequently, we provide guidance on how to collect this evidence. Our focus is on measurement instruments with multiple items; that said, the fundamental principle of searching for validity-supporting evidence applies to all kinds of research instruments, including single-item measurement instruments.

4. Practical Guidance on Assessing Validity of the Measurement Instrument

According to Loevinger (1957), the process of validating a measurement instrument—regardless of whether it is in development or already established—should encompass three phases, as depicted in Figure 2: substantive, structural, and external.

- In the substantive phase, researchers are tasked with grounding their measurement instrument in theory, aligning it with previous literature until their measurement instrument adequately reflects the content of the underlying construct.
- During the structural phase, attention shifts to evaluating the structural and psychometric properties of the measurement instrument, such as item correlations, factor structure, and internal consistency.
- Finally, in the external phase, researchers should check the alignment of their measurement instrument with other criteria and similar tests, ensuring the absence of distortions (Flake, Pek & Hehman 2017; Loevinger, 1957).

The validation process is not necessarily strictly linear, nor is it ever complete. Instead, it is an ongoing process in which researchers collect different types of validity-supporting evidence to build an argument for the validity of their measurement instruments (Flake et al., 2017; Hughes, 2018). In the following, we illustrate different types of validity-supporting evidence and the associated tests using the measurement of political trust as an example. Table 1 in the Appendix summarizes all validity aspects relevant to measurement instruments, including their definitions, examples, empirical tests, and references.

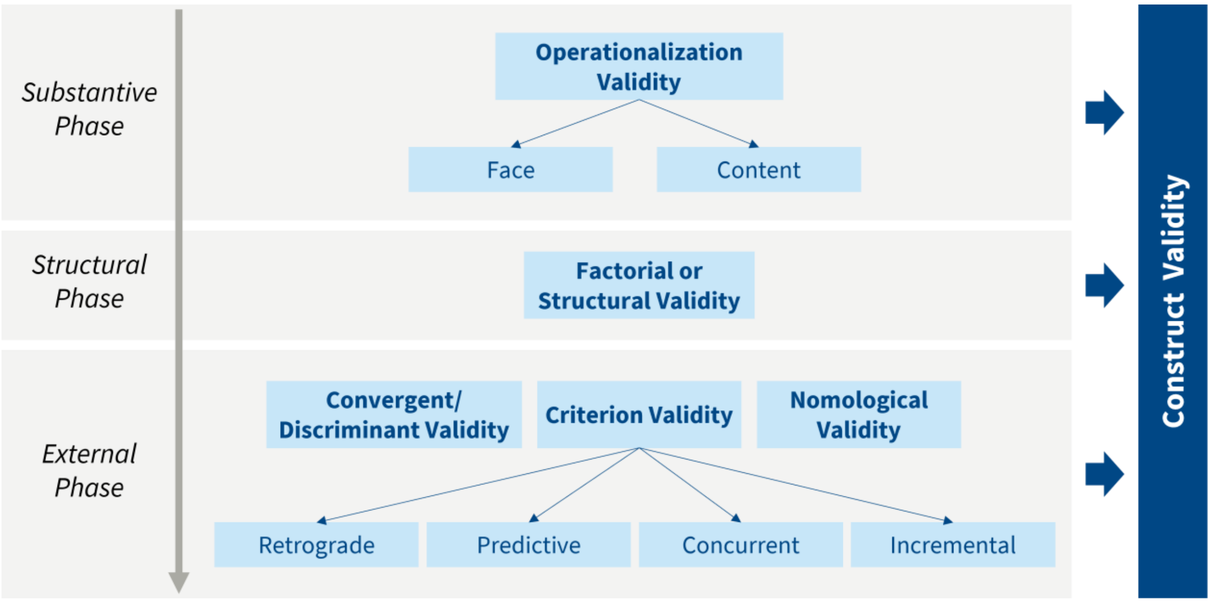

Figure 2: Validation Process Based on Loevinger (1975)

4.1 Substantive Phase

Operationalization Validity. During the substantive phase of validation, researchers must guarantee *operationalization validity* (historically called *translation validity*). It refers to the correct operationalization of a construct (Drost, 2004). Does the operationalization—i.e., the "translation" of the construct into specific indicators, such as survey items—represent the true meaning of the construct? To answer the question of operationalization validity, face and content validity must be assessed.

First, *face validity* is the judgment of whether or not a measurement appears to reflect the construct that is being measured (Holden, 2010). For example, does a measurement of political trust look like it measures political trust? The decision on whether it does is generally based on a qualitative judgment of the researcher or assessed by applying non-expert ratings (e.g., "What do you think this is measuring?"). Thus, tests for face validity usually involve a mix of subjective judgment and critical evaluations. In this procedure, mostly non-experts might be confronted with the measurement instrument (either a single item or multiple items) and asked to evaluate the instrument's appropriateness for measuring the given construct (see Zaichkowsky, 1985). Here, the most popular decision rule is to retain the items and instruments with sufficient agreement that the respective item or instrument is appropriate (for a systematic review of different cutoff strategies, see Hardesty & Bearden, 2004). Beware, even if a measurement instrument is rated positively, this does not guarantee that the underlying concept is well captured. Face validity mostly depends on a subjective and frequently vague evaluation. For that reason, it is often seen only as a starting point in finding validity-supporting evidence (Sartori, 2010).

Second, *content validity* is about how well a measurement instrument covers the range of meanings included within a concept (Nunnally, 1994; Sireci, 1998). For example, does a given measurement instrument cover all aspects of political trust (i.e., trust in national parliament, politicians, and political parties, maybe also trust in international and supranational institutions; Turper & Aarts, 2017) and not just a small part of it (e.g., only trust in national parliaments)? Interestingly, content validity remains one of the most frequently reported dimensions of validity until today, dating back to earlier work by Rulon (1946), Mosier (1947), and Gulliksen (1950a). Evidence on content validity aims to demonstrate a direct and theoretically grounded correspondence between a construct's theoretical and the measurement's actual content (Hughes, 2018).

To ensure content validity during the scale construction process, researchers should first define the domain and the dimensions of the target construct they intend to measure (Gehlbach & Brinkworth, 2011; Hughes, 2018; Nunnally, 1994). The domain refers to the overarching scope encompassed by the construct, while dimensions represent its distinct components. In this process, researchers identify, evaluate, and adapt the most relevant construct definitions, often drawn from authoritative sources such as dictionaries or classification manuals (e.g., American Psychiatric Association, 2013). Additionally, conducting a literature review aids in determining the depth and breadth of the construct, ensuring its appropriate conceptualization within the research field based on the chosen definition (Gehlbach & Brinkworth, 2011). This step lays the foundation for the subsequent development of the item universe, which comprises all potential items that could be included in the measurement instrument.

Next, researchers must generate the questionnaire items (Gehlbach & Brinkworth, 2011; Rammstedt, 2004). The development of these items should be guided by theoretical considerations to ensure their adequacy in representing the defined construct. Here, researchers can draw upon literature reviews, interviews, focus groups, or other qualitative data sources to inform this process (Stewart, Lynn & Mishel, 2010). Finally, the items should undergo formalized rating procedures, where multiple experts assess "the quality of the content based on its relevance, representativeness, specificity, and clarity" (Hughes, 2018, p. 768). For multidimensional and multi-item instruments, competent external coders can be tasked with assigning the items to the hypothesized construct dimensions, a process often referred to as *back translation* (Dawis, 1987).[2] Furthermore, researchers can conduct focus group interviews (Flake et al., 2017), cognitive pretests, or apply webprobing techniques (Lenzner, Hadler & Neuert, 2024) to evaluate the appropriateness of items for a specific research context.

In summary, researchers should apply various checks and methods to ensure that their measurement instruments capture all relevant content dimensions of the target construct. Likewise, researchers who rely on existing measurement instruments and do not create new ones should look at the development process of the instrument they intend to use. This checking process includes reviewing how the item universe was defined, whether and how experts assessed it, and how the final items were selected.

4.2 Structural Phase

Factorial or structural validity. The terms *factorial* and *structural validity* refer to the extent to which the number and nature of a construct's dimensions, as defined by the measurement instrument, match the theorized number and nature of the construct's underlying dimensions, an idea dating back to Loevinger (1957). The question, then, is whether the proposed items of the theoretical concept (e.g., political trust) measure a single underlying construct. In principle, the empirical structure could be unidimensional (i.e., all items in a scale focus on one latent dimension) or multidimensional (i.e., the items in a scale form several independent dimensions; Piedmont, 2014). Testing a measurement instrument's factorial or structural validity requires researchers to apply a combination of empirical methods to demonstrate a correspondence between the theoretically expected and empirically observed dimensions (i.e., factors; Clark & Watson, 2019; Flake et al., 2017).

[2] Back translation is also a (not unproblematic) translation method wherein a translated text is translated back into the source language, allowing for a comparison between the original and the translated text version to assess the accuracy of the translation itself.

To demonstrate a scale's factorial or structural validity, internal consistency, homogeneity, and measurement invariance must be tested. They are all related to and contribute to assessing a measurement instrument's factorial or structural validity. *Internal consistency*, a measure of reliability, typically captures the degree to which the items of a scale are demonstrably correlated with each other. To demonstrate internal consistency, researchers should first start by evaluating the items' response distributions and inter-item correlations (Flake et al., 2017; Piedmont & Hyland, 1993; Stewart et al., 2010). Oftentimes, researchers want to consider eliminating items with highly skewed and unbalanced distributions, as they usually convey very little information and correlate weakly with other items (for a detailed rationale, see Clark & Watson, 2019). Next, researchers can calculate measures of internal consistency, such as Cronbach's alpha (α), tau-equivalent reliability (ρ_T; Cho & Kim, 2015), and McDonald's omega (ω; McDonald, 2013).

For a measurement to be internally consistent for a given context, researchers must also demonstrate the overall reliability of their measurement instrument (i.e., whether the instrument consistently elicits the same results each time it is applied). When only cross-sectional data are available, researchers should rely on the internal consistency measures described above as a remedy to demonstrate the interrelatedness of their empirical measurement scores, such as Cronbach's α and McDonald's ω. Another way to demonstrate reliability is to test the agreement of measures on the same subjects across multiple applications. Test-retest procedures thus provide a more direct way to demonstrate reliability because they might not be affected by systematic measurement errors on a single occasion of data collection (Guttman, 1945; McCrae, Kurtz, Yamagata & Terracciano, 2011).[3]

Once internal consistency is confirmed, researchers need to assess the homogeneity (or unidimensionality) of their measures, which is an aspect of reliability. Homogeneity indicates whether the items of a scale assess one, and only one, underlying latent factor or construct (Briggs & Cheek, 1986; Clark & Watson, 2019). It follows that internal consistency is a necessary but not sufficient condition for homogeneity, as homogeneity can only be demonstrated if all items in a scale are related (Clark & Watson, 2019). To establish the homogeneity of a scale, researchers must show that the scale items adequately measure the underlying factor or construct (McDonald, 1981; Boyle, 1991). This requires statistical analytic tools that provide information on the unidimensionality of the item intercorrelations. Such tools are, for instance, exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and item response theory (IRT; Brahma, 2009; Ziegler & Hagemann, 2015). Adequate formal tests to determine the correct number of dimensions or components are parallel analysis (Horn, 1956), minimum average partial test (Velicer, Eaton & Fava, 2000), or empirical Kaiser criterion (Braeken & Van Assen, 2017).

If the construct of interest is assumed to be multidimensional (i.e., it consists of more than one latent factor), a factor analysis must also be conducted to compare the theorized and empirically observed factor structure. Likewise, hierarchical or multilevel structures must be represented in the empirical factor structure and can be tested using second-order factor analysis (Gould, 2015). To determine the match between the theorized and empirical factor structure, researchers should define several competing CFA models and identify the best-fitting model using standard fit indices (for a more detailed discussion on the selection of fit indices, see Kline, 2014).

Ultimately, researchers should also provide evidence of *measurement invariance* (Leitgöb et al., 2022). That is, whether the empirical relations between the indicators and the latent variables are

[3] When empirical data are not available, instrument developers and instrument users can use the Survey Quality Predictor (SQP) to obtain a quality prediction for their items. SQP is a web-based software freely available at https://sqp.gesis.org/. It is intended to predict the quality of instruments measuring continuous latent variables based on their formal-linguistic characteristics (e.g., the formulation of the item, the characteristics of the response scales).

equivalent across different groups (e.g., gender, educational background, voters vs. non-voters; Hughes, 2018). Measurement invariance is a substantial part of factorial or structural validity as it provides evidence that the identified factor structure is equivalent (and thus comparable) across all groups of interest. Researchers can accomplish this by, for example, using multi-group factor analysis, in which increasingly constrained CFA models across groups are computed and compared (Hirschfeld & Von Brachel, 2014; Tracey & Xu, 2017). In sum, by evaluating internal consistency, homogeneity, and measurement invariance, researchers can provide evidence for a measurement instrument's factorial or structural validity, demonstrating that it accurately captures the intended dimensions or factors of the measured construct.

4.3 External Phase

Convergent and discriminant or divergent validity. Convergent and discriminant (or divergent) validity, two terms coined by Campbell and Fiske (1959), refer to types of validity-supporting evidence in the external phase. On the one hand, convergent validity is the degree to which two measurement instruments of the same or similar construct(s) are related to each other. High correlations of these constructs indicate that there is validity-supporting evidence. What constitutes a sufficiently strong correlation to demonstrate convergent validity remains hotly debated, with scholars applying many different cutoff values. Cohen (2013) provides a rough guideline for interpreting correlation coefficients as small (.10), medium (.30), and large (.50). However, the interpretation of associations between two variables should always include a critical reflection of the research design, incorporating theoretical considerations and related research findings.

Discriminant or divergent validity, on the other hand, is the extent to which the scores of a measurement instrument are not or only very slightly correlated with theoretically different but close concepts. Ideally, there are no relationships between unrelated concepts. For instance, one would expect that a measurement of political trust should correspond closely to a similar measurement of trust in the core political institutions (convergent validity). In contrast, one would expect only a low correlation with a measurement of political knowledge (discriminant or divergent validity). Establishing discriminant or divergent validity is thus a way of showing that the measurement of interest differs from other measurement instruments and, therefore, is not redundant.

Researchers can inspect the correlations between related and unrelated constructs to check for convergent and discriminant validity. To do this systematically, they can perform a multitrait-multimethod (MTMM) matrix analysis (Campbell & Fiske, 1959). An MTMM matrix consists of the traits under consideration (i.e., the theoretical constructs) and their different measurement methods (e.g., different data types; Dumenci, 2000). For example, consider three different traits related to political trust: trust in government, trust in regional government, and trust in local government. Suppose these traits are measured using three different methods: two self-report questionnaires using different Likert-type scales and an external assessment. The resulting matrix consists of nine correlations related to validity (3 traits x 3 methods). Each cell in the matrix represents the correlation between two measures, either of the same trait using different methods (i.e., mono-trait-heteromethod) or of different traits using the same or different methods (i.e., heterotrait-monomethod and heterotrait-heteromethod). Monotrait-heteromethod correlations are used to assess convergent validity. These correlations describe the relationship between two measures of the same trait using two different methods. High correlations here indicate that different methods are indeed measuring the same trait, thus confirming convergent validity.

However, the MTMM matrix provides much more information than just the monotrait-heteromethod entries. It allows for a comprehensive evaluation of both convergent and discriminant validity through relative comparisons. Researchers can analyze heterotrait-monomethod and heterotrait-heteromethod correlations to gain insights into the relationships between different traits and their measurement methods. For instance, high correlations between different measures of the same trait (monotrait-heteromethod) indicate convergent validity, whereas low correlations between different traits measured with multiple methods (heterotrait-heteromethod) indicate discriminant validity. By comparing the relative sizes of these correlations, researchers can assess whether the methods used to measure different traits are appropriately capturing the distinctiveness of each construct.

In summary, MTMM matrices are a powerful tool providing a wealth of information about convergent and discriminant validity. By examining and comparing the various types of correlations within the matrix, researchers can gain deep insights into the validity of their measurement methods and the relationships between different constructs.

Nomological validity and nomological nets (or networks). The term ‘nomological' is derived from Greek and means ‘lawful.' Hence, *nomological validity* is the extent to which a construct behaves according to the hypothesized relationships with other variables. This idea was first introduced by Cronbach and Meehl (1955). Researchers can assess nomological validity by developing a nomological net using theory and common sense. A *nomological net* represents the concepts or constructs involved in a study, their observable manifestations, their interrelationships (i.e., empirical relations), and their relationships to other theoretical constructs (Clark & Watson, 2019; Cronbach & Meehl, 1955; Preckel & Brunner, 2020). For example, a nomological net for political trust could include political engagement, satisfaction with the government, voting behavior, political sophistication, political interest, and education.

To test nomological validity, researchers must evaluate the observed nomological net in terms of the theoretically expected relations (Cronbach & Meehl, 1955; Podsakoff & MacKenzie, 1994). Hagger, Gucciardi & Chatzisarantis (2017), among others, provide a detailed set of four relevant testing steps: First, in the specification step, researchers must focus on the “core” components of the theory and establish a priori the relationships between the proposed components and hypothesized effects based on theoretical considerations. These components and their relationships are the minimum required to test whether the theory is supported. Second, in the investigation step, researchers must identify specific tests to confirm or reject the specified network based on observations and data. This process includes all steps to maximize the quality of evidence for evaluating the relationships. For example, researchers must consider the appropriateness of the research design and analysis techniques (such as structural equation modeling) and the sample size with its statistical power. Third, in the interpretation step, researchers must interpret the test results and decide on the appropriateness of the construct. Here, it is crucial to avoid post-hoc adjustments that counter the previously assumed relations between the components. Fourth, in the replication or reformulation step, researchers should replicate or reformulate the effects and nomological relationships to verify the effects and provide robust evidence of nomological validity (Lindsay, 2015).

Criterion or criterion-related validity: Retrograde, predictive, concurrent, and incremental validity. *Criterion* or *criterion-related validity* refers to the ability of a measurement instrument to predict an independent criterion. For example, a measure of political trust that one seeks to validate should be able to predict a person's voting behavior (the criterion) such that people with higher levels of political trust are likely to show higher voter turnout and opt for centrist parties (Hooghe, 2017). The criterion is a benchmark or a gold standard against which the measure is tested.

Depending on whether the criterion was measured before, at the same time, or after the instrument whose validity is to be established, criterion validity can be divided into retrograde (or postdictive or retrospective) validity, concurrent validity, and predictive validity, respectively. While *retrograde* (or *postdictive* or *retrospective*) *validity* examines the relationship between the measurement and a criterion in the past (e.g., a measurement for political trust and actual voting behavior in the last election), *predictive validity* examines whether a measurement accurately predicts a behavior on a criterion measured in the future. The latter is often seen as the gold standard in psychological testing (e.g., university admission or employee selection). *Concurrent validity* is like predictive validity but simultaneously assesses the relationship between the measure and the criterion. Essentially, this means that the new measurement instrument and the established measurement instrument are given to the same group of people, and the researcher assesses whether the new instrument produces results consistent with the established instrument. Another criterion-related validity is *incremental validity*. It assesses whether a new measurement instrument has greater predictive power than an already established instrument (Sechrest, 1963). The main question is whether a measurement instrument helps predict a variable beyond what can already be predicted by other measures.

To gather criterion-related evidence (retrograde, predictive, concurrent, and incremental validity), researchers usually inspect the bivariate correlations between the measurement of their construct and selected criterion variables. Whenever researchers want to control for background variables (e.g., age and gender), they can apply more advanced regression methods (Shou et al., 2022). Importantly, the quality of criterion-related evidence depends on the correct choice of a gold standard criterion. Therefore, tests of criterion validity should also include evidence that the putative gold standard is an accurate estimate of the underlying construct's true value (Bellamy et al., 2006).

5. Conclusion and Recommendations

Across the past decades, the concept of validity has undergone several shifts and redefinitions in social and behavioral science research. However, the fundamental question that defines validity—whether the research design and the measurement instruments are true to what they are supposed to represent or capture—has remained unchanged. All existing validity aspects relate to the quality and trustworthiness of the research, either in terms of the research design or the measurement. In survey research, establishing the validity of measurement instruments is essential, no matter whether one or multiple manifest indicators measure an underlying latent construct. To demonstrate measurement validity, researchers must provide various kinds of validity-supporting evidence that show the strength of their theoretical and methodological foundations. This guideline outlines the most fundamental types of validity-supporting evidence and ways to test for them. It also stresses that validation efforts depend on the research context and method applied.

As we highlighted, researchers should view validation as an ongoing process and not a single outcome of the research process itself (Strauss & Smith, 2009; Flake et al., 2017). Ergo, they should carefully consider all relevant validity aspects to enhance the quality and trustworthiness of their research results. They can do this in many ways, depending on whether they want to create new measurement instruments, implement existing measurement instruments in their study, or analyze already existing data. Researchers developing new instruments should be guided by the current state of research and should not underestimate the extensive effort required to collect all kinds of validity-supporting evidence. Researchers who "borrow" existing instruments could evaluate an instrument's documentation[4] and check previous validation

[4] A database where you can find such documentation for tested German, English, and multilingual measurements instruments is the open access repository for measurement instruments called ZIS. It is freely accessible at https://zis.gesis.org/en.

efforts in light of the intended application field before applying a specific instrument, rather than simply believing an instrument is valid because it has been used in previous studies. Likewise, a research design might work in one context but not necessarily in another. Applying it in a non-suitable context would likely bias the analysis and lead to misleading or incorrect conclusions. In general, a lack of conceptual clarity is likely to affect all facets of validity negatively. Therefore, researchers should establish and draw from explicit construct definitions, which provide answers to the most relevant characteristics of the construct (Podsakoff et al., 2016). Relying on established data quality frameworks, such as the Total Error Framework for survey data (Biemer, 2010), can guide researchers in their documentation (see Daikeler et al., 2023).

We encourage researchers in all disciplines to address validity questions in the research design and measurement stages while considering their specific research context. Although validation efforts may seem tedious and time-consuming, they are paramount to any research process. Apart from being necessary, they ultimately provide researchers with a critical mindset and the ability to produce high-quality scientific work.

6. References

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3), 529–546.

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959. https://doi.org/10.1177/001316448004000419

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014*). Standards for educational and psycho-logical testing.* American Educational Research Association. https://www.apa.org/science/programs/testing/standards

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (Vol. 5). American psychiatric association Washington, DC.

Aronson, E., Ellsworth, P. C., Carlsmith, J. M., & Gonzales, M. H. (1990). Methods of Research in Social Psychology (2nd ed.). McGraw-Hill.

Aronson, E., Wilson, T. D., & Brewer, M. B. (1998). Experimentation in social psychology. In The Handbook of Social Psychology, 1, 99.

Bellamy, N. (2014). Principles of clinical outcome assessment.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5), 817-848.

Birkenmaier, L., Lechner, C., & Wagner, C. (2023). The search for solid ground in text as data: A systematic review of validation practices and practical recommendations for validation. *Communication Methods and Measures*, 1–29. https://doi.org/10.1080/19312458.2023.2285765

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

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences*, 12(3), 291-294.

Braeken, J., & Van Assen, M. A. (2017). An empirical Kaiser criterion. *Psychological Methods*, 22(3), 450-466.

Brahma, S. S. (2009). Assessment of construct validity in management research: A structured guideline. *Journal of Management Research*, 9(2), 59-71.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, 54(1), 106–148.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105. https://doi.org/10.1037/h0046016

Campbell, D. T., & Stanley, J. C. (2015). Experimental and quasi-experimental designs for research. Ravenio books.

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology*, 97(1), 186–202. https://doi.org/10.1037/a0015618

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, 18(2), 207–230. https://doi.org/10.1177/1094428114555994

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12), 1412–1427. https://doi.org/10.1037/pas0000626

Cohen, R. J., & Swerdlik, M. E. (2005). Psychological testing and assessment: An introduction to tests and measurement (6th ed.). McGraw-Hill.

Cook, T. D., & Campbell, D. T. (1979). Quasi-Experimentation: Design and Analysis Issues for Field Settings. Houghton Mifflin.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281.

Cureton, E. (1951). Validity. In Educational Measurement. American Council on Education.

Daikeler, J., Fröhling, L., Sen, I., Birkenmaier, L., Gummer, T., Schwalbach, J., Silber, H., Weiss, B., Weller, K. & Lechner, C. (2024). Assessing data quality in the age of digital social research: A systematic review. *Social Science Computer Review*, 08944393241245395.

Danner, D. (2016). Reliability – The precision of a measurement. In GESIS Survey Guidelines (Issue December). https://doi.org/10.15465/gesis-sg_en_011

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology*, 34(4), 481–489. https://doi.org/10.1037/0022-0167.34.4.481

Drost, E. A. (2004). Validity and reliability in social science research. *Education Research and Perspectives*, 38(1), 105–125.

Dumenci, L. (2000). Multitrait-multimethod analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of Applied Multivariate Statistics and Mathematical Modeling* (pp. 583–611). Aca-demic Press. https://doi.org/10.1016/B978-012691360-6/50021-5

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: current practice and recommendations. *Social Psychological and Personality Science*, 8(4), 370–378. https://doi.org/10.1177/1948550617693063

Garb, H. N. (1984). The incremental validity of information used in personality assessment. *Clinical Psychology Review*, 4(6), 641–655. https://doi.org/10.1016/0272-7358(84)90010-2

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology*, 15(4), 380–387. https://doi.org/10.1037/a0025704

Goodwin, L. (2002). Changing conceptions of measurement validity: An update on the new standards. *Journal of Nursing Education*, 41 (3), 100-106.

Gould, S. J. (2015). Second order confirmatory factor analysis: An example. In J. M. Hawes & G. B. Glisan (Eds.), *Proceedings of the 1987 Academy of Marketing Science (AMS) Annual Confer-ence* (pp. 488–490). Springer International Publishing. https://doi.org/10.1007/978-3-319-17052-7_100

Gregory, R. J. (2004). *Psychological testing: History, principles, and applications*. Allyn & Bacon.

Gulliksen, H. (1950a). Intrinsic validity. *American Psychologist*, 5(10), 511–517. https://doi.org/10.1037/h0054604

Gulliksen, H. (1950b). *Theory of mental tests.* Wiley.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4), 255–282.

Hagger, M. S., Gucciardi, D. F., & Chatzisarantis, N. L. D. (2017). On nomological validity and auxiliary assumptions: The importance of simultaneously testing effects in social cognitive theo-ries applied to health behavior and some guidelines. *Frontiers in Psychology*, 8, 292317. https://www.frontiersin.org/article/10.3389/fpsyg.2017.01933

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of Business Research*, 57(2), 98–107.

Harkness, J. A., Edwards, B., Hansen, S. E., Miller, D. R., & Villar, A. (2010). Designing questionnaires for multipopulation research. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 33–57). Wiley.

Hirschfeld, G., & Von Brachel, R. (2014). Improving Multiple-Group confirmatory factor analysis in R–A tutorial in measurement invariance with continuous and ordinal indicators. *Practical As-sessment, Research, and Evaluation*, 19(1), 7.

Holden, R. R. (2010). Face validity. In *The Corsini Encyclopedia of Psychology* (4th ed., pp. 637–638). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470479216.corpsy0341

Hooghe, M. (2017). Trust and elections (E. M. Uslaner, Ed.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190274801.013.17

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179-185.

Hughes, D. (2018). Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures. In *The Wiley handbook of psychometric testing: A multidisciplinary ap-proach to survey, scale and test development.* John Wiley & Sons Ltd.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment*, 15(4), 446–455. https://doi.org/10.1037/1040-3590.15.4.446

Jiménez-Buedo, M., & Miller, L. M. (2010). Why a trade-off? The relationship between the external and internal validity of experiments. THEORIA. *Revista de Teoría, Historia y Fundamentos de La Ciencia*, 25(3), 301–321.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–342.

Kelley, T. L. (1927). *Interpretation of educational measurements.* World Book Company.

Kline, P. (2014). *An easy guide to factor analysis.* Routledge.

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., De Roover, K., ... & van de Schoot, R. (2022). Measurement invariance in the social sciences: Historical development, methodologi-cal challenges, state of the art, and future perspectives. *Social Science Research*, 102805.

Lenzner, T., Hadler, P., & Neuert, C. (2024, forthcoming). *Cognitive Pretesting*. GESIS Survey Guidelines.

Lin, W.-L., & Yao, G. (2014). Concurrent validity. In *Encyclopedia of Quality of Life and Well-Being Research* (pp. 1184–1185). Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_516

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, 26(12), 1827–1832. https://doi.org/10.1177/0956797615616374

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research,* 35(6), 382-386

Markus, K. A. (2008). Constructs, Concepts and the Worlds of Possibility: Connecting the Measurement, Manipulation, and Meaning of Variables. *Measurement: Interdisciplinary Research & Perspective*, 6(1–2), 54–77. https://doi.org/10.1080/15366360802035513

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psycholo-gy Review*, 15(1), 28–50.

McDermott, R. (2011). Internal and external validity. In James N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge Handbook of Experimental Political Science*, 27.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 34(1), 100–117.

McDonald, R. P. (2013). *Test theory: A unified treatment.* Psychology Press.

McIntire, S. A., & Miller, L. A. (2010). *Foundations of psychological testing* (3rd ed.). Sage.

Messick, S. (1994). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *ETS Research Report Series*, 1994(2), i–28. https://doi.org/10.1002/j.2333-8504.1994.tb01618.x

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement*, 7(2), 191–205. https://doi.org/10.1177/001316444700700201

Murphy, K. R., & Davidshofer, C. O. (1998). *Psychological testing: Principles and applications* (4th ed.). Prentice-Hall. https://doi.org/10.1201/9781315380797

Muthén, L. K., & Muthén, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 9(310), 599–620.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research & Perspective*, 10(1–2), 1–29.

Newton, P. E., & Baird, J. A. (2016). The great validity debate. *Assessment in education: principles, policy & practice*, 23(2), 173-177.

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational & psychological assessment*. Sage.

Nunnally, J. C. (1994). Psychometric theory 3E. Tata McGraw-Hill Education.

Peeters, M. J., & Harpe, S. E. (2020). Updating conceptions of validity and reliability. *Research in Social and Administrative* Pharmacy, 16(8), 1127-1130.

Piedmont, R. L. (2014). Factorial validity. In Encyclopedia of Quality of Life and Well-Being Research (pp. 2148–2149). Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_984

Piedmont, R. L., & Hyland, M. E. (1993). Inter-item correlation frequency distribution analysis: A method for evaluating scale dimensionality. *Educational and Psychological Measurement*, 53(2), 369–378. https://doi.org/10.1177/0013164493053002006

Podsakoff, P. M., & MacKenzie, S. B. (1994). An examination of the psychometric properties and nomological validity of some revised and reduced substitutes for leadership scales. *Journal of Applied Psychology*, 79(5), 702.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2016). Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences. *Organizational Research Methods*, 19(2), 159–203. https://doi.org/10.1177/1094428115624965

Preckel, F., & Brunner, M. (2020). Nomological nets. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences*. Springer International Publishing. https://doi.org/10.1007/978-3-319-28099-8

Rammstedt, B. (2004). *Zur Bestimmung der Güte von Multi-Item-Skalen: Eine Einführung*. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-201443

Ruch, G. M. (1924). *The improvement of the written examination*. Scott, Foresman.

Rulon, P. J. (1946). *On the validity of educational tests.* Harvard Educational Review, 16, 290–296.

Sartori, R. (2010). Face validity in personality tests: Psychometric instruments and projective techniques in comparison. *Quality & Quantity*, 44(4), 749–759.

Schneider, I. (2017). Can we trust measures of political trust? Assessing measurement equivalence in diverse regime types. *Social Indicators Research*, 133(3), 963–984.

Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement*, XXIII(1), 153–158.

Shou, Y., Sellbom, M., & Chen, H.-F. (2022). Fundamentals of measurement in clinical psychology. In G. J. G. Asmundson (Ed.), *Comprehensive Clinical Psychology* (2nd Edition) (pp. 13–35). Elsevier. https://doi.org/10.1016/B978-0-12-818697-8.00110-2

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, 45, 83–117. https://doi.org/10.1023/a:1006985528729

Slack, M. K., & Draugalis, J. R. (2001). Establishing the internal and external validity of experi-mental studies. *American Journal of Health-System Pharmacy*, 58(22), 2173–2181. https://doi.org/10.1093/ajhp/58.22.2173

Stewart, J. L., Lynn, M. R., & Mishel, M. H. (2005). Evaluating content validity for children's self-report instruments using children as content experts. *Nursing Research*, 54(6), 414–418.

Stewart, J. L., Lynn, M. R., & Mishel, M. H. (2010). Psychometric evaluation of a new instrument to measure uncertainty in children and adolescents with cancer. *Nursing Research*, 59(2), 119–126. https://doi.org/10.1097/NNR.0b013e3181d1a8d5

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, 5(1), 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Tracey, T. J., & Xu, H. (2017). Use of multi-group confirmatory factor analysis in examining measurement invariance in counseling psychology research. *The European Journal of Counselling Psychology*, 6(1), 75–82.

Turper, S., & Aarts, K. (2017). Political trust and sophistication: Taking measurement seriously. *Social Indicators Research*, 130(1), 415–434. https://doi.org/10.1007/s11205-015-1182-4

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy, 41-71.

Wang, Y. A., & Eastwick, P. W. (2020). Solutions to the problems of incremental validity testing in relationship science. *Personal Relationships*, 27(1), 156–175. https://doi.org/10.1111/pere.12309

Wang, Y. A., & Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*, 4(1), 1–17. https://doi.org/10.1177/25152459209

Ward, C., & Kennedy, A. (1994). Acculturation strategies, psychological adjustment, and sociocultural competence during cross-cultural transitions. *International Journal of Intercultural Relations*, 18(3), 329–343.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, 11(3), 1–22. https://doi.org/10.1371/journal.pone.0152719

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79–94. https://doi.org/10.20982/tqmp.09.2.p079

Zaichkowsky. (1985). Measuring the involvement construct. *Journal of Consumer Research*, 12(3), 341–352.

Ziegler, M., Booth, T., & Bensch, D. (2013). Getting entangled in the nomological net: Thoughts on validity and conceptual overlap. *European Journal of Psychological Assessment*, 29(3), 157–161. https://doi.org/10.1027/1015-5759/a000173

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items. *European Journal of Psychological Assessment.*

7. Appendix

Table 1: Overview of Validation Aspects for Measurement Instruments

		Definition	Example	Exemplary Empirical Tests	Key Literature
	Construct validity				
		= encompasses all kinds of validity-supporting evidence relevant to the interpretation or the meaning of respondents' answers to survey questions (broad definition)	A measure of political trust does not only look like it measures political trust but actually does measure political trust.	Methods that check for the different validity aspects (see below)	Cronbach & Meehl (1955); Loevinger (1957); Campbell & Fiske (1959); Messick (1994); Strauss & Smith (2009)
SUBSTANTIVEE PHASE	*Operationalization (or translation) validity*				
		= extent to which a construct is truthfully operationalized	----	Assessment via face and content validity	Drost (2004); Clark & Watson (2019); Podsakoff et al. (2016
		Face validity			
		= content of the measure appears to reflect the construct being measured	A measure of political trust looks like it measures political trust and not something else (e.g., political engagement).	Qualitative methods (e.g., judgment); non-expert ratings; probing questions; examining interrater reliability	Mosier (1947); Holden (2010)
		Content validity			
		= extent to which an instrument covers the range of meanings included within a construct that is being measured	A measure of political trust should cover all aspects of political trust (e.g., trust in national parliament, politicians, and political parties) and not just a small part of it (e.g., only trust in politicians).	Expert panels or judges; checking the item development process; literature review; cognitive pretesting and web-probing (Lenzner et al., 2015)	Rulon (1946); Mosier (1947); Gulliksen (1950a, 1950b); Aiken (1980); Sireci (1998)

		Definition	Example	Exemplary Empirical Tests	Key Literature
	Factorial (or structural) validity				
STRUCTURAL PHASE		= extent to which the number and nature of a construct's dimensions as defined by the instrument correspond to the theorized number and nature of the construct's underlying dimensions	Political trust (conceptualized as trust in political institutions) is a one-dimensional latent concept. So, a person's score on a political trust scale would reflect only trust in political institutions and not trust in the economy.	Evaluating the items' response distributions and inter-item correlations; confirmatory approaches to determine the degree of "fit" between expected and obtained structure when using an already validated scale: confirmatory factor analysis (CFA), structural equation modeling (SEM), item response theory (IRT); tests for measurement invariance; when developing a new scale: exploratory factor analysis (Yong & Pearce, 2013); makes only sense for multiple-item measures	Cohen & Swerdlik (2005); Piedemont (2014)
	Convergent validity				
EXTERNAL PHASE		= degree to which two measurements of a construct or similar constructs are related	A measure of political trust should be related to similar constructs, such as political sophistication.	Nomological network as a guiding framework; correlation analysis (+/- high); multi-trait multi-method	Campbell & Fiske (1959); Ziegler, Booth, & Bush (2013)
	Discriminant (or divergent) validity				
		= degree to which two similar constructs are distinct	A measure of political trust should not relate too much to stranger-face trust.	Nomological network as a guiding framework; correlation analysis (+/- low)	Campbell & Fiske (1959); Ziegler, Booth, & Bush (2013)
	Criterion (or criterion-related) validity				
		= extent to which a construct correlates with external criteria (i.e., established measures that have shown to be valid)	-----	Correlation analysis; multivariate regression methods	Cohen & Swerdlik (2005)
		Retrograde (or postdictive or retrospective) validity			
		= is one approach of criterion validity that examines the relationship between the measure and a criterion in the past	There is a relationship between political trust (present) and actual voting behavior in the last election (past).	Correlation analysis; multivariate regression methods	

		Definition	Example	Exemplary Empirical Tests	Key Literature
EXTERNAL PHASE					
		Predictive validity			
		= is one approach of criterion validity that estimates how accurately a measurement predicts the performance of a criterion measured at a time in the future	Political trust (present) predicts whether a person will vote in the next election (future).	Correlation analysis; multivariate regression methods	Lin & Yao (2014); McIntire & Miller (2010)
		Concurrent validity			
		= is one approach of criterion validity that estimates the relationship between the measure and the criterion simultaneously	There is a strong relationship between political trust (present) and political engagement (present).	Correlation analysis; multivariate regression methods	Gregory (2004); McIntire & Miller (2010); Murphy & Davidshofer (1998); Lin & Yao (2014)
		Incremental validity			
		= improvement obtained in the predictive power of a new instrument compared to an already established instrument	A measure of political trust that consists of multiple items (e.g., trust in the national parliament, politicians, and political parties) performs better than a single-item measure (e.g., trust in the government).	Hierarchical multiple regression analysis (change in adjusted R^2); structural equation modeling possible, but recommended to conduct a power analysis (e.g., by using Monte Carlo simulations; see Muthén & Muthén, 2002), or the Shiny app pwrSEM by Wang & Rhemtulla (2021)	Sechrest (1963); Garb (1984); Hunsley & Meyer (2003); Wang & Eastwick (2020); Westfall & Yarkoni (2016)
	Nomological validity				
		= extent to which a construct behaves according to the hypothesized relationships with other constructs/concepts/variables	A nomological net of political trust could include political engagement, satisfaction with the government, voting behavior, political sophistication, political interest, and education	Nomological net(work)s	Cronbach & Meehl (1955); Preckel & Brunner (2020); Hagger et al. (2017)