Multiple Historic Trajectories Generate Multiplicity in the Concept of Validity

Yingying Han
Institute for Science in Society, Radboud University, The Netherlands

Although researchers agree on the importance of validity, they have not yet reached a consensus on what validity consists of. This article traces the historic trajectories of validity theory development in three research traditions: psychometrics, experiment in social settings, and animal models of psychiatric disorders, showing that the multiplicity in the concept of validity is shaped by its multiple historic trajectories and reflects the diversity of practices and concerns in different research traditions. I argue that specifying validity of *what target practice and* for *what purpose in discussions helps to connect validity to its rich context that gives rise to its specific meaning and relevance.*

1. Introduction

Validity, understood as the quality of being sound, well-founded, or accepted, holds a central position in both social science and biomedical research methodologies. The recent concerns surrounding the "replication crisis" or "reproducibility crisis" brought to the forefront several issues linked to validity (for an overview, see Fidler and Wilcox 2021). In these discussions, validity is often perceived as a unified concept. However, the term's technical interpretations differ across disciplines, leading to a plethora of validity-related concepts and an expansive taxonomy. Notably,

I would like to thank Henk de Regt, Willem Halffman, and Luca Consoli for developing the manuscript's concept through brainstorming, and for their inspiring discussions, comments, and feedback that have significantly shaped it. My gratitude goes to Uljana Feest and Maarten Derksen for their invaluable feedback on an earlier draft. I also appreciate the Institute for Science in Society's Research Quality Team and the Research Therapy group meeting led by Lara Keuck and Alfred Freeborn, for their constructive discussions and feedback. Further, I am thankful to the anonymous reviewers for their constructive criticism and insightful suggestions, and to the editor for his support.

Perspectives on Science 2024, vol. 32, no. 4

https://doi.org/10.1162/posc_a_00624

within the realm of educational and psychological assessment alone, over 150 types of validity have been identified[1] (Newton and Shaw 2014), many of which have been widely used in methodological discourses. This article aims to clarify the differences among some of these validity taxonomies by making explicit the context in which they were used to show that they addressed different concerns. The discussion will focus on some validity concepts that have been recently discussed in philosophy of science. The contextual aspect will be substantiated by historical narratives of the theoretical development of these validity taxonomies in order to advocate for a contextually sensitive approach to the concept, linking it to its practical tradition and purpose of use.

The discourse in philosophy of science currently gravitates towards three types of validity: construct, internal, and external. Construct validity, predominantly situated within the framework of psychological testing, has been analyzed extensively with references to the psychometrics tradition (Stone 2019; Feest 2020). External validity has been discussed in other contexts such as experimental economics (Guala 1999, 2005) and animal and social experimentation[2] (Steel 2004, 2007; Guala 2010), which ultimately fitted in the internal-external validity dichotomic framework first proposed by Donald Campbell in the experimental psychology tradition (Campbell 1957; Campbell and Stanley 1963). Another salient thread in the ongoing philosophical dialogue pertains to the validity of randomized control trials (RCTs). While Nancy Cartwright (2009, 2011) raised concerns about the external validity of RCTs without referencing Campbell's seminal framework, it was the work of Jiménez-Buedo and Russo (2021) who reintegrated Campbell's validity taxonomy into discussions. They proposed to break the internal-external validity dichotomy and "re-embed" construct validity in the framework, highlighting the intricacies of Campbell's validity theory. This work illustrated the value of rekindling validity concepts with their historical and contextual background.

However, the philosophical discussions frequently divorce validity concepts from their rich historical context and the original research traditions from which they emerged. Such oversights risk stripping away the nuanced technicalities, spawning potential conceptual ambiguities. With

1. These kinds or types of validity were identified by the modifier added to validity, such as "external validity," "construct validity," and so on. Newton and Shaw analyzed the 151 validities they found and reported that "Somewhere in the region of 28 are mere synonyms of others in the list. Although quite a few of the terms have been given different meanings by different authors, there is actually quite a lot of overlap in the meanings associated with most of the terms" (Newton and Shaw 2014, p. 6).

2. In these contexts, the term "extrapolation" was often used instead of external validity.

this paper, I aim to bridge this gap, starting with an attempt to disentangle the complex web of validity taxonomies, tracing their unique historical trajectories and conceptual origins. By closely examining these taxonomies through the lens of the traditions they originated from and the practices they were applied to, I aspire to showcase the importance of context in comprehending the multiplicity of validity. Central to this contextual approach are two guiding questions: validity *of* which research practice and validity *for* what purpose?

Specifically, validity "of" paints a clearer picture of the research tradition in question, anchoring discussions to specific communities and their associated activities in the research practice; validity "for" sheds light on the epistemic and pragmatic objectives the practice is tailored for. Together, these delineations illuminate the nuances shaping the validity concepts and their intricate interplay. Employing this dual framework of validity "of" and validity "for" paves the way for a more coherent discussion. Connecting the ongoing discourse with the theoretical underpinnings of relevant research traditions forestalls conceptual ambiguities and the pitfalls of redundant exploration.

Guided by this framework, the study offers a deep dive into three primary traditions: the psychometrics tradition (section 2), underscoring construct validity (Cronbach and Meehl 1955); the experimental psychology tradition (section 3) featuring internal and external validity introduced by Campbell (Campbell 1957; Campbell and Stanley 1963); and the biomedical research tradition, especially animal modeling of psychiatric disorders (section 4), mainly focusing on Paul Willner's validity taxonomy (i.e., face, predictive, and construct validity; Willner 1984, 1986). Spanning sections 2 through 4, the discourse on validity will be systematically contextualized, aligning with the research practices in focus (i.e., validity *of*) and the objectives these validity concepts serve (i.e., validity *for*). Concluding section 5 revisits the central thesis: rooting validity concepts in their native context accentuates their inherent significance and relevance.

By proposing the validity "of" and validity "for" framework, this paper aspires to equip philosophers, methodologists, and researchers with a calibrated lens to navigate the intricate terrains of validity.

2. Validity of Testing: The Psychometrics Tradition

The technical concept of validity has a significant history in the field of psychometrics, tracing back throughout the twentieth century and continuing to be a subject of ongoing debate. Comprehensive historical analyses of validity theory in education and psychological testing have been provided by various scholars such as Newton and Shaw (2014) and Slaney (2017). Drawing on their work, this section does not aim to

reiterate their detailed reviews, but rather to emphasize a crucial aspect: the development of the concept of validity is deeply intertwined with its historical context. This becomes particularly evident when comparing validity taxonomies from diverse historical contexts, especially when centered on their respective research practices and objectives.

2.1. Validity Emerged as an Assessment of Measurement Instruments

The term "validity" first emerged in the technical context of the psychometrics tradition in 1921, as delineated in the "Report of the Standardization Committee" (Buckingham et al. 1921):

> Two of the most important types of problems in measurement are those connected with the determination of what a test measures, and of how consistently it measures. The first should be called the problem of validity, the second, the problem of reliability. (Buckingham et al. 1921, p. 80)

From its outset, validity was identified as a significant measurement challenge for scholars. To understand this concept fully, it is useful to examine its historical context, specifically the measurement challenges faced at the time, who was addressing them, and their approaches.

In the late nineteenth century, several countries in North America and Europe began revising their assessment procedures in education. There was an increased reliance on exams to evaluate educational outcomes, influencing decisions about students, educators, and institutions. Key researchers such as James McKeen Cattell, Herbert Rice, and Edward Thorndike led the "measurement movement," advocating for tests that could gauge school performance, predict educational success, and support teaching methods (Newton and Shaw 2014). These tests quickly became popular among educators. However, the rapid growth of educational testing brought about criticisms. Some scholars questioned the traditional essay tests, believing they were too reliant on the evaluator's subjective judgment. This led to a move towards standardized tests with formats like true/false or multiple-choice questions, which were seen as "more objective and impartial" (Ruch 1929, p. 22).

Meanwhile, psychologists sought to measure abilities like intelligence, which had relevance beyond just school. The Binet-Simon tests, early forms of intelligence tests, were introduced in France in 1905. These tests were later adapted in various countries, including the USA, England, Italy, and Germany. A notable version is the American adaptation by Yerkes and his team, used for evaluating US Army recruits during World War I. Despite criticisms of its validity by later researchers, it was widely used

during that period (Carson 1993). The success of these tests encouraged the wide employment of group tests. After the war, psychological testing for vocational guidance became more common globally, with many countries establishing institutes that used psychologists for vocational testing and advice (Burt 1924).

With the increasing use of psychological testing, both psychologists and personnel administrators shared educators' concerns and sought more objective and standardized assessment methods. This led to the development of validity as a specific technical term. Organizations like the American Psychological Association (APA) and what would later become the American Educational Research Association (AERA) formed committees to standardize testing methods. Within the AERA's efforts to address standardization, the term 'validity' was specifically introduced in the "Report of the Standardization Committee" (Buckingham et al. 1921). Both educational and psychological associations formed committees to address this issue, leading to the concept of validity that we understand today. This theoretical development will be further discussed in sections 2.2–2.4.

Truman Lee Kelley, a member of this committee, provided a straightforward definition of the "classic conception of validity" (Hood 2009, p. 453):

> The problem of validity is that of whether a test really measures what it purports to measure, while the question of reliability is that of how accurately a test measures the thing which it does measure. (Kelley 1927, p. 14)

Kelley's definition underscored two pivotal aspects. Firstly, validity was posited as a singular, unmodified concept, distinct from later multi-faceted interpretations. Secondly, it emphasized gauging the degree to which a measurement tool, whether exams or psychological assessments, captured its intended purpose.

2.2. The "Trinitarian" Conception of Validity: Epistemic Beliefs and Professional Practices

The term "validity" refers to the degree to which a test accurately measures what it purports to measure. Determining this led to a common practice in the 1920s of comparing test results with external benchmarks, either through logical reasoning or empirical data collection and statistical analysis (Newton and Shaw 2014; Slaney 2017).

In education, analyzing the content of exams by logical reasoning was a popular method to ensure alignment with educational objectives. Face validity was proposed, leaning on the intuitive or a "common-sense relationship" of the test content that "appear[s] on their face" (Mosier 1947,

p. 192). However, content validity soon emerged to assess the true relevance of test items. For ability tests, the empirical route was favored, utilizing correlation coefficients to compare test scores with the intended criteria. This quantitative approach, while useful, occasionally led to an overreliance on statistical correlations without due consideration of the actual content relevance (Newton and Shaw 2014).

To standardize testing practices, including establishing criteria for validity, the APA initiated a committee led by Lee Cronbach. Their efforts culminated in the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (*Standards* 1st ed.; American Psychological Association, American Educational Research Association, and National Council on Measurements Used in Education 1954).[3] This foundational text and its successors have since guided both academic discourse and professional application. The evolution of the validity concept and its classifications within the *Standards* will be further discussed in subsections 2.3 and 2.4.

Standards 1st ed. (American Psychological Association, American Educational Research Association, and National Council on Measurements Used in Education 1954) defined validity as "the degree to which the test is capable of achieving certain aims" (1954, p. 13). This notion emphasized the aims of tests, and validity was explicitly specified in degrees. In a departure from earlier singular definitions of validity, this edition introduced four types of validity: content, predictive, concurrent, and construct validity. This classification addressed the diverse needs of various test developers and users. For instance, content validity was associated with educational testing, "evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn" (1954, p. 13). It was noted to be "especially important in the case of achievement and proficiency measures" (1954, p. 13). Predictive validity was grounded in the ability-testing tradition, defined as "showing how well predictions made from the test are confirmed by evidence gathered at some subsequent time" (1954, p. 13). Concurrent validity, akin to predictive validity, gauged the correspondence of test scores with concurrent (rather than future) criteria, often utilized in clinical or vocational settings.

Construct validity, "investigating what psychological qualities a test measures" (1954, p. 14), was a novel concept first introduced in the

3. The 2nd edition published in 1966 changed its title to *Standards for Educational and Psychological Tests and Manuals*. The 3rd edition published in 1974 again modified the title to *Standards for Educational and Psychological Tests* and in the 4th edition, published in 1985, the title was changed to *Standards for Educational and Psychological Testing*. This title has been kept in the 5th edition published in 1999 and the 6th edition published in 2014. I thus refer to different editions as "*Standards* nth ed." for short in this article.

Standards 1st ed. and subsequently elaborated in an influential paper by Cronbach and Meehl (1955). As revealed by Cronbach and Meehl (1955), this validity type was devised for tests like personality assessments, which lacked clear comparison criteria. To address this, Campbell and Fiske (1959) introduced the multitrait-multimethod (MTMM) methodology, emphasizing higher inter-correlations for tests measuring the same construct than those measuring different constructs.

Standards 2nd ed. (American Psychological Association, American Educational Research Association, and National Council on Measurement in Education 1966) merged the similar predictive and concurrent validity into one category, termed "criterion-related validity," which compared test scores with "one or more external variables considered to provide a direct measure of the characteristic or behavior in question" (1966, p. 12). This three-type classification of validity (content, criterion-related, and construct) was retained in *Standards* 3rd ed. (American Psychological Association, American Educational Research Association, and National Council on Measurement in Education 1974) and has since become the dominant framework in psychometrics.

Although the *Standards* did not explicitly promote the fragmentation of validity, the three-type taxonomy has become one of the most recognized and utilized in psychometrics and related fields. This three-fold classification, sometimes referred to as the "(un)Holy Trinity" or "Trinitarian" doctrines of validity (Guion 1980; Slaney 2017), underscored the diverse focus areas in different testing scenarios: content validity for educational contexts, criterion validity for vocational guidance, and construct validity for personality evaluations.

2.3. Validity (Re)Unified as Construct Validity

During the 1960s and 1970s, there was a trend of selectively applying the authoritative *Standards*. Many opted for the straightforward use of one type of validity for a specific test, sidestepping the more labor-intensive process of considering all three or four validity types as suggested (Newton and Shaw 2014). Recognizing the limitations of such a fragmented approach to validity, several scholars aimed for a more integrated perspective, with Samuel Messick's methodology emerging as particularly impactful.

Messick drew significant inspiration from Cronbach's insights on validity. He adopted the notion of construct validity from Cronbach and Meehl (1955), positioning it at the heart of his validity theory. Building upon Cronbach's (1971) idea that "one validates, not a test, but an interpretation of data arising from a specified procedure" (1971, p. 447), Messick emphasized that the crux of validity was the interpretation of test outcomes, not

the tests *per se*. He conceptualized construct validity as "the evidential basis for inferring a measure's meaning" (Messick 1975, p. 955).

The *Standards* 4th ed. (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 1985) integrated elements of Messick's validity theory. This edition reconceptualized the prior three validity types as three "forms of evidence"—content, criterion-related, and construct, aligning with Messick's perspective on validity as the evidential basis for drawing inferences (1975, 1980). In this version, validity was distinctly framed as a "unitary concept" (1985, p. 9), described as "[t]he degree to which a certain inference from a test is *appropriate or meaningful*" (1985, p. 94; my emphasis), echoing Messick's stance.

By the time of *Standards* 5th ed. (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 1999), the vestiges of "Trinitarian" validity had largely faded. Instead of three forms of evidence, five evidence sources were introduced: test content, response processes, internal structure, relation to other variables, and testing consequences. What had previously been seen as evidence for different types of test validity was then seen as evidence for the validity of a construct. This shift solidified the integrated view of validity in academic discussions by the close of the twentieth century. However, its practical ramifications and influence are less clear. Although methodologists seemed to agree upon a unified concept of validity under construct validity as suggested by the latest versions of the *Standards*, diverse validity types are still referenced across textbooks, publications, and various academic and philosophical discourses.

2.4. Recent Developments

Recent discussions around the concept of validity within the psychometric tradition have primarily revolved around the nature, significance, and scope of construct validity (hence the one and only validity).

Michael T. Kane (1992) introduced an argument-based perspective on validity -> validity:

> Validity is associated with the interpretation assigned to test scores rather than with the test scores or the test. The interpretation involves an argument leading from the scores to score-based statements or decisions, and the validity of the interpretation depends on the plausibility of this interpretive argument. (Kane 1992, p. 527)

This perspective by Kane was incorporated into both the *Standards* 5th ed. (American Educational Research Association, American Psychological

Association, and National Council on Measurement in Education 1999) and 6th ed. (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). Both editions articulated the validation process as formulating a "validity argument," described as "[a]n explicit scientific justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) of test scores for their intended uses" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 1999, p. 184; the underlined part was added (without underlining) in American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014, p. 225).

Challenging the perspectives of both Kane and Messick on construct validity as an attribute of measurement results[4], Borsboom and collaegues (2004, 2009) advocated a return to the foundational definition, emphasizing validity as an intrinsic quality of the measurement instrument:

> Construct validity theory holds that (a) validity is a property of test score interpretations in terms of constructs that (b) reflects the strength of the evidence for these interpretations … We propose an alternative view that holds that (a) validity is a property of measurement instruments that (b) codes whether these instruments are sensitive to variation in a targeted attribute. (Borsboom et al. 2009, p. 135)

Hood (2009) proposed that the perspectives of Borsboom and Messick on (construct) validity could be seen in tandem. While Borsboom primarily delved into the ontological aspects of constructs (i.e., whether a construct must be real for the test to be valid), Messick focused on epistemological aspects of measurement (i.e., how to assess interpretive inferences).

The evolution of validity theory within the psychometric domain remains a dynamic and ongoing endeavor.

2.5. Contextualization: Validity *Of* Psychometric Testing *For* Measurement and Its Interpretation

Within the psychometric tradition, testing serves as the cornerstone methodology, encompassing the activities of test creation, application, and evaluation. Consequently, the concept of validity primarily pertains to the

4. There has also been some discussion about whether psychological constructs are measurable. For instance, see Michell (2013).

domain of testing within this tradition. This focus persists, even as the emphasis has shifted from the mere measurement aspect of testing to encompass its interpretation and appropriate utilization. The notion of validity itself emerged in the 1920s in response to the need for standardized educational and psychological tests. The diverse applications of testing, ranging from assessments of achievement and aptitude to personality evaluations, gave rise to multiple types of validity in the 1950s. These classifications endured, despite subsequent efforts to consolidate them under the umbrella of construct validity, whether through the presentation of varied validity evidence in the 1980s or the development of diverse validity arguments in recent years. The evolving classifications of validity across historical phases not only signify the maturation of our theoretical understanding of assessment but also mirror the evolving demands of testing methodologies. The validity *of* psychometric testing thus encompasses a distinctive array of validity taxonomies intrinsically intertwined with the historical trajectory of this research tradition.

This pragmatic nature of validity has been evident within the psychometrics tradition since its inception. Initially, validity was employed to assess the extent to which a test measures "what it claims to measure." This directly relates to the conceptual understanding of the targeted construct and the underlying purpose of the measurement. Recognizing the infeasibility of a one-size-fits-all approach, psychometricians introduced various types of validity "[t]o determine how suitable a test is for each of [the different] uses" (*Standards* 1st ed., p. 13). For instance, the "Trinitarian doctrine" determined assessments for diverse testing objectives: 1) content validity for evaluating capabilities, 2) predictive validity for practical forecasting, and 3) construct validity for theory development. As the scope of validity expanded to encompass not only measurement but also its theoretical interpretation and application, construct validity emerged as the singular, comprehensive validity type in psychometrics due to its wider evaluative capabilities. Validity *for* measurement or its interpretation underscores the pragmatic characteristics of psychometric testing, highlighting the relevance of specific taxonomies within their practical contexts.

3. Validity of Experimentation: The Campbellian Tradition

Researchers within the same discipline, such as psychology, often employ varied research practices. While testing and experimentation often go hand in hand, the validity paradigms crafted for psychometrics were not adequately equipped for experimental requirements. Recognizing this gap, Donald Campbell, in the 1950s, put forth a tailored validity framework for experimentation. This section elucidates the historical progression of

the Campbellian validity taxonomy, its practical applications in psychology and other experimental research fields, and contrasts it against the validity taxonomies employed in psychometric testing, highlighting that specifying validity "of" and "for" helps to disambiguate similar validity concepts used in different research contexts.

3.1. Emergence and Context of Campbellian Taxonomy amidst Experimental Advancements

Experimental psychology witnessed a significant shift with the integration of Ronald Fisher's statistical methodologies during the 1930s, culminating in a broader embrace of his experimental designs by the 1950s (Gigerenzer et al. 1989). Fisher's contributions, encompassing the principles of randomization, the analysis of variance (ANOVA), and the null hypothesis, became foundational in experimental design and analysis. Yet, as Cronbach (1957) insightfully remarked, "Fisher made the experimentalist an expert puppeteer, able to keep untangled the strands to half-a-dozen independent variables" (1957, p. 675).

However, the very zeal for the Fisherian methodology also sparked a mix of skepticism and critique among statisticians and psychologists alike. Despite acknowledging Fisher's undeniable impact, Cronbach pointed out that "[h]is sophistication in data analysis has not been matched by sophistication in theory" after praising Fisher's contribution (Cronbach 1957, p. 675). Central to the reservations about the Fisherian methodology was their agricultural origin, specifically designed to assess fertilizer effects on crops. Such origin posed stark contrasts to the nuanced requirements of psychological experimentation, particularly in controlling and measuring social variables (Campbell and Stanley 1963; Cronbach 1975; Meehl 1978; Campbell 1986). For instance, while observing crops might remain largely unaltered by an observer's presence, the mere act of observation could significantly skew human behavior (Landsberger 1957). Moreover, while Fisher's null hypothesis significance testing effectively identified interventions' impacts, its capacity for rigorous theory testing was contentious.

These reservations resonated with Campbell, who perceived an overreliance on Fisher's randomized models as "overwhelming" and having "an implicit, complacent assumption that meticulous care in this regard took care of all experimental validity problems" (Campbell 1986, p. 68). Such concerns motivated Campbell to introduce a new validity taxonomy to help rectify Fisher's prevailing "lopsided and complacent emphasis" on the randomized design as the "only methodological control that needed to be taught" (1986, p. 68). This framework, fundamentally different from the validity taxonomies in psychometrics, focused predominantly on

internal and external validity[5]: Internal validity addressed the question "did in fact the experimental stimulus make some significant difference in this specific instance?" and external validity asked, "to what populations, settings, and variables can this effect be generalized?" (Campbell 1957, p. 297; Campbell and Stanley 1963, p. 5). Although the link to causation was not yet explicitly stated, the concept of internal validity can be interpreted as an assessment of the causal relation between a particular experimental manipulation and its effect. If an experiment was internally valid, external validity came into play to determine to what target the (causal) effect could be generalized. This interpretation logically implies that external validity depends on internal validity.

This paper later expanded into a chapter in *Handbook of Research on Teaching* (Gage 1963) and was subsequently republished as a standalone book (Campbell and Stanley 1963). At the outset of this influential chapter, Campbell clarified his deviation from the Fisherian experimental tradition, articulating:

> It is not a chapter on experimental design in the Fisher tradition, in which an experimenter having complete mastery can schedule treatments and measurements for optimal statistical efficiency with the complexity of design emerging only from that goal of efficiency. Insofar as the designs discussed in the present chapter become complex, it is because of the intransigency of the environment: because, that is, of the experimenter's lack of complete control. (Campbell and Stanley 1963, p. 1; original emphasis)

They underscored their work as a "strong advocacy of experimentation" (Campbell and Stanley 1963, p. 2) with a caveat—"on more pessimistic grounds—not as a panacea, but rather as the only available route to cumulative progress." (1963, p. 3). The Campbellian validity taxonomy's theoretical and pragmatic underpinnings were reflected in two aspects. First, it aimed to navigate challenges inherent in Fisher's randomization-centric approach, emphasizing factors potentially undermining validity. To address the practical intricacies of managing uncontrollable factors in social experiments, Campbell delineated twelve "factors jeopardizing the validity of various experimental designs" and argued that "[f]undamental to this listing is a distinction between *internal validity* and *external validity*"

5. The internal/external validity taxonomy was not the only taxonomy of validity proposed by Campbell. As mentioned in section 2.2, he also introduced the taxonomy of convergent validity and discriminative validity along with the MTMM method to evaluate construct validity in psychometrics. Campbell kept his discussion about the validity of testing "quite distinct" from his discussion about the validity of experiments (Newton and Shaw 2014, p. 4).

(Campbell and Stanley 1963, p. 5; original emphases). Second, the work's focus was not to cultivate validity theory but to guide experimental design, especially in contexts where randomized design might be impractical, making them vulnerable to validity threats. For example, Collier's (1944) classroom study illustrated how significant events (e.g., France's WWII fall) could confound social experiments. This exemplified Campbell's taxonomy's inception as a response to the unique epistemological and pragmatic demands of social and behavioral experimentation.

3.2. Refinement of the Campbellian Taxonomy

Following the establishment of the internal-external validity framework, questions arose about the precise boundaries between these concepts. Notably, Campbell himself construed internal validity in a more restrictive sense than commonly assumed. While internal validity assessed if the experimental manipulation created a difference, it remained non-specific about which factors were responsible for that difference in a given experiment. For example, if one accounts for the placebo effect (where outcomes arise from participant expectations rather than direct experimental influence), does this enhance internal or external validity? Campbell argued it augmented external validity, as it pinpoints which factor (be it a confounding element or a deliberate manipulation) resulted in the outcome. Reflecting on the ambiguity, he noted that "half of my own students fail to answer [this] question correctly" (Campbell 1986, p. 67). Additionally, Campbell highlighted a recurring misinterpretation:

> [B]oth those who have enthusiastically adopted the distinction and those who oppose it, have most frequently redefined it to epitomize all the differences between pure laboratory experimentation and field tryouts of ameliorative programs. (Campbell 1986, p. 67)

To redress these conceptual nuances and criticisms, Campbell refined his taxonomy, further categorizing the types of validity (Cook and Campbell 1979). Under "Internal validity," two categories were delineated: "statistical conclusion validity" and refined "internal validity." The former dealt with "sources of random error and with the appropriate use of statistics and statistical tests" while the latter explored "systematic bias" (Cook and Campbell 1979, p. 80). Meanwhile, in the chapter on "External validity", the concept of "external validity" was sharpened, and "construct validity" was incorporated. Drawing from psychometrics, construct validity addressed generalizing from specific experimental contexts to broader theoretical frameworks, while the refined external validity considered generalizations "to specific populations of persons, settings, and times that have a grounded existence" (Cook and Campbell 1979, p. 82). Although

construct validity is found in both psychometrics and Campbellian traditions, their connotations and emphases differed. Within psychometrics, construct validity eventually became the essential and overarching concept, but within the experimental practices, it served as a facet of external validity. Given its absence in the original Campbellian taxonomy, construct validity often remained overshadowed by the internal-external validity dyad in methodological debates on social science experimentation.

Depending on the experimental objective, Campbell proposed varying hierarchies of validity. In theory-focused experiments, the "priority ordering" was "internal, construct, statistical conclusion, and external validity." For applied endeavors, the priority hierarchy shifted to "internal validity, external validity, construct validity of the effect, statistical conclusion, and construct validity of the cause." Additionally, Campbell accentuated the primacy of internal validity, emphasizing its fundamental importance "for both basic and applied researchers" (Cook and Campbell 1979, p. 83). The validity of the causal effect within a particular experimental setting (i.e., internal validity) is fundamental and a prerequisite for further generalization or extrapolation (i.e., external validity or construct validity). Although the generalization or extrapolation from experimental observations to the underlying constructs is also important, without ensuring the establishment of a causal effect in the first place (i.e., internal validity), there would be nothing to generalize or extrapolate. Therefore, internal validity must be prioritized in Campbell's view. The marked emphasis on internal validity diverges from the psychometrics tradition and underscores the main goal of experimental research: establishing causal relations. This highlights the need to frame validity with respect to the specific research practice in question (i.e., validity "of") and the purpose that underpins such practice (i.e., validity "for").

3.3. Campbellian Taxonomy's Ripple Effect on Economic Experimentation

The Campbellian framework has garnered widespread recognition beyond psychology, finding traction across diverse fields that employ experimental methods. For instance, this taxonomy of validity is found in the *Cambridge Handbook of Experimental Political Science* (Druckman et al. 2011). When educators were surveyed on "What should graduate students know about research and statistics?," "Internal and external validity of designs" emerged as a frontrunner, surpassing 30 other topics covering research process and design, data gathering, and statistics (Mundfrom et al. 1998). Remarkably, despite the initial reluctance, the concepts of internal and external validity "eventually became in economics the strict twin concepts

by which the meaningfulness of experiments was assessed" (Heukelom 2011, p. 24).

This extended reach of the Campbellian framework attests to its "pandisciplinary" influence (Kazdin 2000). However, the universality of the terms sometimes clouded their nuanced application within specific contexts. This is illustrated vividly in the historical trajectory of the Campbellian validity framework within experimental and behavioral economics.

In the 1950s, Vernon Smith laid the groundwork for experimental economics. As this innovative approach emerged, discussions abounded about the "proper method of conducting experiments." Yet, the lexicon of validity, as conceptualized by psychologists, remained sidelined. Heukelom suggested this might have been due to concerns that "the psychologists' way of dealing with validity risked creating a division between an inside world of the laboratory and an outside 'real' world of the economy and its actors. Such an interpretation, they argued, would be directly against experimental economists' conception of experiments in economics" (2011, p. 20).

When experimental economics began incorporating psychological notions of validity, it did so tentatively, with some economists suggesting that efforts might be better redirected elsewhere, rather than tangling with the intricacies of external validity (Brookshire et al. 1987). The introduction of internal validity followed, albeit with "reservations with regard to the new methodological concepts" (Heukelom 2011, p. 21).

However, by the late 1990s, Campbell's distinctions had firmly embedded themselves in both experimental and behavioral economics, a sentiment reflected by Heukelom's description of them as "household concepts." Although the 1980s and 1990s saw the taxonomy either reinterpreted or labeled as a borrowed concept "from the psychological literature," the narrative shifted in the late 1990s and early 2000s (Heukelom 2011). Francesco Guala emerged as a pivotal figure, foregrounding validity as a methodological cornerstone of economic experiments (Guala 1999, 2005). Guala emphasized the relation of internal validity with "some particular aspect of a laboratory system" in contrast with external validity, which is linked to the "natural settings" that are "outside the experimental circumstances" (Guala 2005, p. 142). He also tied these considerations to debates within philosophy of experimentation (e.g., Mayo 2008).

Guala's definitions, while echoing the Campbellian tradition, showcased two deviations. Firstly, the boundary between internal and external was different. While Guala drew the line between a controlled environment (i.e., the lab) and the natural settings (i.e., the real world), Campbell's distinction lay between one particular experimental setting versus other settings which can be in or outside a lab as long as some element of the

setting is changed. The scope of Campbell's internal validity was much smaller than that of Guala's notion. This reflected the different experimental practices and their aims. Campbell aimed to "extend the epistemology of the experimental method into nonlaboratory social science" and focused on "applied social science, on treating the ameliorative efforts of government as field experiments" (Campbell 1997, p. 36). In this sense, Campbell positions experimentation in the field from the very beginning, rather than considering it a test case in the laboratory to be generalized later to the field. In contrast, Guala argued that "laboratory experimentation is supposed to provide more efficacious tools to tackle the Duhem-Quine problem" (Guala 2005, p. 61), which considered experimentation as a tool to test economic theories in a controlled environment such as laboratories to avoid confounding factors in the field. Here experiments started in the lab and the conclusions from experiments must be generalizable to the field, making it necessary to consider validity as internal or external to the experimental environment, namely the laboratory.

Secondly, when considering the trade-off, Guala leaned more towards external validity[6], echoing the sentiments of economists that the ultimate fruits of their labor lay beyond the confines of the lab. In contrast, Campbell's emphasis hinged more on internal validity, reflecting his advocacy for the significant epistemological role of field experimentation. These variances underscore the multifaceted interpretations of validity contingent on the distinct objectives and contexts of social psychology and economics. This emphasizes the importance of delineating validity in terms of specific practices (validity "of") and purposes (validity "for") to cater to the diverse contexts in which it is applied.

3.4. Contextualization: Validity *Of* Experimentation *For* Causal Inferences in Social Settings

Understanding validity requires a precise contextual lens. While one might be inclined to classify validity frameworks by historical timelines or disciplinary boundaries, this method proves inadequate in some cases. Notably, within psychology, varying validity taxonomies can exist simultaneously. Even though both psychometric testing and psychological experimentation stem from the same academic lineage, they have birthed distinct validity taxonomies. In light of this, I propose to frame validity in terms of its associated research practice (validity "of") and its overarching goal (validity "for").

6. Interestingly, Guala's discussion on external validity in philosophy of science rippled beyond the realm of economics or validity of experiments, touching upon the evaluation of animal models in biomedical research (Guala 2010).

Within the psychometric tradition, the focus of validity gravitates towards assessing the fidelity of measurement tools or interpreting their outputs. Conversely, in the broader sphere of social sciences, the essence of experimental validity evaluates the robustness of causal inferences derived from experiments. Though the validity concepts from different research practices may sometimes share terminological similarities, their meanings can diverge significantly based on their application. For instance, while "construct validity" in testing might signify the relation between a measurement (or its interpretation) and its theoretical correlate, in an experimental setting, it bridges specific experimental outcomes with broader theoretical frameworks. By aligning discussions of validity with their specific research practices, we can avoid potential ambiguities. Hence, pinpointing the validity "of" a particular practice becomes a valuable anchor, grounding debates within the rich context of methodological and theoretical evolution.

Moreover, pinpointing validity "for" a specific purpose is invaluable, especially when considering that a single validity taxonomy can be used across varied research endeavors with diverse objectives. Consider the Campbellian validity taxonomy, tailored for experimentation. It has been embraced across fields like psychology, economics, and political science. While these disciplines share a common research practice (i.e., performing experiments) and similar applications (i.e., developing theories of causal relations and/or interventions based on these theories), the diverse purposes of experiment in these fields are reflected in the varied conceptions of validity.

In advocating for an "experimenting society," Campbell emphasized the role of experimental studies in generating robust evidence for policy decisions (Campbell 1973, 1991; Dehue 2001). He constructed a comprehensive framework to address challenges associated with both internal and external validity, especially in non-randomized experiments. For Campbell, the delineation between internal and external validity was not simply the division between the laboratory and the outside world as his social experiments inherently took place within real-world contexts. Conversely, economists treat experiments as controlled simulations, from which they can deduce insights applicable to real-world economic scenarios (Guala 2005). This epistemological divergence between Campbell's and Guala's views led to subtle shifts in their conceptualizations regarding the nature of external validity and its applicability. This issue has also been extensively discussed by Jiménez-Buedo and colleagues in the contexts of economics (Jiménez-Buedo and Miller 2010; Jiménez-Buedo 2011) and clinical research (Jiménez-Buedo and Russo 2021). The varying visions and objectives of experiments across these traditions lead to subtle, yet

significant, differences in understanding validity. Clarifying the research practice (validity "of") and its associated objectives (validity "for") aids in elucidating the underlying assumptions, interpretations, challenges, and the multifaceted nature of the concept.

4. Validity of Animal Models for Human Psychiatric Disorders

In philosophy of science, external validity commonly falls under extrapolation discussions (Guala 2010; Reiss 2019), emphasizing the application of causal inferences from experiments to broader contexts (Steel 2007, 2010). While Guala's extrapolation leans towards "real world" economic activities, in biomedical research, the focal point centers around human physiological, behavioral, or pathological processes. However, due to technical and ethical barriers, human experimentation is often neither feasible nor appropriate, giving rise to the longstanding practice of using non-human animal models. To critically assess these models, particularly for human psychopathology, intricate validity taxonomies play a significant role.

4.1. The Classic Validity Framework in Animal Research: Willner's Taxonomy

In the realm of human psychiatric disorders, animal models frequently prioritize "observable behavioral changes," especially when comprehensive biological insights into the disorder remain elusive. Interpretation of behavioral changes, inherently subjective, pose challenges for objective evaluation. Willner observed "[w]ith the eclipse of behaviorism, animal models accompanied animal learning theory into a period of decline, which in both cases lasted until the late 1960s" (Willner 1986, p. 678). The "turning point of animal models" was attributed to McKinney and Bunney's (1969) introduction of "minimal requirements" for an animal model of depression. Willner interpreted their work as aiming:

> to introduce a much needed objectivity into the study of animal models: there should be observable behavioural changes which can be objectively evaluated, independent observers should agree on objective criteria for drawing conclusions about subjective state, and the system should be reproducible by other investigators. (Willner 1986, p. 678)

In this context of "much needed objectivity" upon the "rapidly expanded" studies of animal models, resembling the historical context in which validity in psychometrics was introduced, Willner designed a taxonomy grounded on McKinney and Bunney's seminal work. His taxonomy comprising face validity, predictive validity, and construct validity

(Willner 1984, 1986), was reminiscent of psychometric terminology (see section 2). Much like Campbell and Guala in their respective domains, Willner repurposed these psychometric concepts, tailoring them to fit the intricacies of animal models:

> face validity: the phenomenological similarities between the model and the condition being modelled … predictive validity concerns the success of predictions made from the model, and construct validity concerns its theoretical rationale. (Willner 1984, p. 1)

Willner expressly acknowledged that the concept of construct validity was "a term borrowed from psychological testing," citing Vernon (1963) as a reference. Within Vernon's comprehensive study, he dedicates an entire chapter to the intricacies of validity. This includes detailed discussions on "face and context validity," "external validation (concurrent and predictive)[7]" and an extended exploration of construct validity, resembling the "Trinitarian" framework established by the *Standards*. Intriguingly, while Vernon's discourse touched upon all three validity types present in Willner's validity taxonomy, Willner cited Vernon only when introducing the notion of construct validity. Of particular note is Vernon's discussion about the controversial nature of face validity, describing it as a concept "which psychometrists unanimously condemn" (Vernon 1963, p. 213).

Rather than discarding face validity due to its contentious reputation in its original domain, Willner recalibrated its application to suit the evaluation of animal models. Within the psychometric tradition, face validity assesses "what a test measures by what it 'looks like', or by analyzing its content subjectively" (Vernon 1963, p. 213). This overt subjectivity appeared to diverge from Willner's pursuit of "objective criteria for drawing conclusions." Due to "uncertainty" in practice, Willner further excluded the more objective etiology and biochemistry criteria from McKinney and Bunney's (1969) ground rules, leaving only the "similarity of symptomatology and treatment as usable criteria for face validity" (Willner 1986, p. 681). Focusing on the behavioral resemblance might be a compromise resulting from the poor understanding of the etiology and biochemical mechanism of psychiatric disorders in human patients at the time and hence the lack of targets to model in animals. In essence, Willner's adaptation of face validity emerged as a thoughtful reconciliation of epistemological concerns and the pragmatic challenges of conceptualizing and evaluating animal models.

7. In this context, "external" refers to "external criterion" as in "criterion-related validity" in the psychometrics tradition, not in the Campbellian sense of "external validity."

As for predictive validity, Willner drew upon Russel's presentation "Extrapolation from Animal to Man" (1964). This talk reiterated the intertwinement between validity of animal models and validity in the other two traditions: predictive validity echoed the same term in the psychometrics tradition and extrapolation was connected to Campbellian external validity. In his talk, Russell highlighted the "measuring" property of animal models used in behavioral tests and characterized them as measuring instruments. This stance was congruent with Willner's approach which championed validity in terms of measurement rather than mere experimentation. Nonetheless, the diverse scholarly interpretations of animal models reflect the evolving and multifaceted nature of validity theories in this domain.

4.2. Recent Developments

Over time, Willner's framework has become a cornerstone in the study of animal models for human psychiatric disorders. His 1984 publication, where he first delineated the validity taxonomies for assessing animal models of depression, has garnered significant academic attention, as evidenced by its numerous citations[8]. Notably, this foundational work has become a touchstone for many subsequent scholars, with many adapting or expanding upon Willner's taxonomy (for a review, see Belzung and Lemoine 2011), reminiscent of the evolving nature of validity theories in psychometrics and social experimentation.

Some scholars refined the classic taxonomy by "hierarchizing" the validity criteria. They either weighted the relative importance or relevance of different validity types (e.g., see Belzung and Lemoine 2011) or introduced a step-by-step evaluation procedure, which resembled two basic tendencies in the psychometric tradition: to make construct validity central (e.g., Messick's attempt of unification) and to schematize the whole of validity (e.g., Kane's argumentation approach). For instance, Van der Staay and colleagues (2009) introduced an "iterative model building" procedure incorporating various validity criteria into the model evaluation stage. Models lacking face validity, homology, or reliability reached a "dead end" in the model evaluation flow chart whereas models that met these requirements proceeded to be assessed in terms of relevance using criteria including generalizability, predictive validity, and construct validity. This type of approach could help address the issue that "[v]alidity is too often asserted in published papers rather than systematically discussed in terms

8. Willner's 1984 paper has been cited 919 times in January 2022 according to Web of Science and 1621 times according to Google Scholar.

of strengths and weaknesses" (Nestler and Hyman 2010, p. 1162). This shift towards a more process-oriented view of validation echoed contemporary debates around construct validity in psychometrics (Alexandrova and Haybron 2016; Feest 2020).

A notable trend in contemporary validity taxonomies for animal models is the incorporation of experimental validities (i.e., internal and external validity) into the classic Willner's framework. This highlights the dual role animal models play: they act both as experimental apparatuses and measurement tools. On the one hand, an animal model is part of an experiment aiming for causal inferences. They are typically developed as alternatives to human subjects in experiments that either present technical challenges or raise ethical concerns (Baetu 2016). Here, the focus on validity pivots towards the theories associated with experimentation. On the other hand, the behaviors exhibited by animals within the experimental setting ought to be authentic measures or indicators of the attributes in question. Consequently, taxonomies rooted in psychometric validity can be fine-tuned to evaluate this aspect of animal models. Recognizing this inherent duality, it becomes apparent that a holistic evaluation requires the incorporation of validity theories from both experimental and measurement traditions.

Ankeny and colleagues' case study on animal models for alcoholism further illustrates the dual nature of animal models (2014). The initial stages in the development of these models focused on selecting organisms based on physiological compatibility and experimental convenience. Subsequently, deliberations emerged "about what constitutes a representative experimental set-up for human behavior." In evaluating models for alcoholism, two perspectives were adopted: viewing them as analogous instruments measuring human conditions, and as experimental designs to investigate specific human behaviors. The "touchstone" criteria for assessing model validity in this field, delineated as Cicero's criteria, encapsulate aspects of face validity (e.g., "animals had to self-administer alcohol by the oral route"), construct validity (e.g., "alcohol should be consumed for its pharmacological properties and not for its taste or caloric properties") and predictive validity (e.g., "tolerance and dependence must emerge, as a result, measured by reduced effects of alcohol consumption and acute withdrawal symptoms") in Willner's framework. While the primary focus of these criteria hinged on the congruence or homology between the animal model and the human situation, the evaluation of the experimental methodology was deemed integral to the overall assessment process.

A feature observed in the initial development phases of animal models for alcoholism was the diversity in the choice of organisms and experimental parameters. However, over time, a trend towards greater standardization

became evident. While this push towards standardization is not exclusive to this field, its implications on the external validity or generalizability of findings have been a point of extensive debate. The push for standardization in animal models encapsulates the dynamic interplay between internal and external validity, a relationship deeply explored within the Campbellian framework (Sullivan 2009; Jiménez-Buedo and Russo 2021; Feest 2022). Though standardized experimental conditions often facilitate greater control, the specificity of these standardized models may restrict the broader applicability of the findings. Consequently, while standardization might strengthen internal validity, it could simultaneously compromise external validity. The optimal balance between these two remains elusive, but the importance of addressing this balance during model evaluations cannot be overstated.

Henderson and colleagues' (2013) systematic review of guidelines for preclinical animal experiments adds another layer to this discussion. Their categorization of recommendations based on threats to internal, construct, and external validity drew parallels with Campbell's taxonomy. Notably, their analysis revealed a predominant focus in these guidelines on internal and construct validity, with external validity, particularly replications, being somewhat underrepresented. Such findings accentuate the important role of experimental validity in both the formulation and evaluation of animal models in biomedical research.

This trajectory of advancements in animal model validity theory reflects a synthesis of insights from both psychometrics and experimental traditions. Employing animals as analogs for humans introduces intricate challenges centered around establishing congruence between the model and its human counterpart. When evaluating animal models for psychiatric disorders, parallels are drawn based on symptoms (face validity), treatment outcomes (predictive validity), and underlying biological processes (construct validity). Yet, when these models transition into the experimental domain, the evaluative scope expands to encompass both dimensions of measurement and broader experimental considerations. The selection between evaluative framework, whether aligned with Willner's or Campbell's perspectives, largely depends on the specific research goals and the envisaged usage of the animal model.

4.3. Contextualization: Validity *Of* Animal Models *For* Human Psychiatric Disorders

Animal models designed for human psychiatric disorders intriguingly meld elements of both measurement and experimentation. On one side, these models utilize animals as indicators, aiming to mirror the extent of human psychiatric pathologies. In such contexts, evaluating the model's

validity necessitates a framework akin to Willner's taxonomy, which is rooted in the psychometrics tradition. Yet, on the other side, these models function within experimental paradigms, serving to decipher causal connections concerning both mechanisms and therapeutic approaches. Here, the Campbellian framework, emphasizing the validity of experimentation, is of paramount importance for critically appraising such endeavors. Additionally, the primary objective of these models is to facilitate extrapolation from animals to humans, inherently linking them with external validity, thus augmenting their experimental dimensions. Consequently, validity of animal models showcases the intricate relations of various validity taxonomies in different traditions and underscores the importance of specifying the validity *of* which practice in both theoretical discourses and practical implementation.

When considering researchers' objectives, two different aspects of external validity become particularly relevant. Some researchers aspire to extrapolate experimental findings from the controlled confines of a laboratory to the more complex dynamics of real-world contexts, echoing Campbell's and Guala's perspectives. In contrast, others seek to apply the causal relations established from an experimented cohort to a distinct target, such as translating insights from animals to humans. The validity of animal models for human psychiatric disorders illustrates this latter dimension of external validity. The emphasis on the "for" aspect of validity is also reflected in policies. For instance, according to the European regulation for "the protection of animals used for scientific purposes" (Directive 2010/63/EU), "it is essential, both on moral and scientific grounds, to ensure that each use of an animal is carefully evaluated as to the scientific or educational validity, usefulness, and relevance of the expected result *of that use*" (Recital 39; emphasis added).

5. Conclusion

Validity, while foundational to research methodology, manifests differently across diverse disciplines and fields. Rather than addressing validity in a general sense, a more nuanced approach involves examining its relevance and implications within specific practices and for particular objectives. In this light, I propose to contextualize validity by specifying validity *of* which practice and *for* what purpose. Validity "of" identifies the practice in question, placing the concept within its distinct disciplinary and historical backdrop. This contextuality illuminates the theoretical underpinnings and practical consequences associated with the term. The orientation towards research practice echoes the recent "practice turn" observed in both sociology of science and philosophy of science (Chang 2004; Russo 2022). Validity "for" foregrounds the pragmatic nexus between validity

theory and real-world validation practices, underscoring the concept's practical essence. This purpose-oriented perspective resonates with ongoing discussions in philosophy of science concerning model evaluation, as exemplified by Parker (2020). By addressing both dimensions—validity "of" and "for"—we recognize the multifaceted nature of the term, tracing its historical lineage, clarifying its diverse uses and meanings, and engaging the recent debates of various branches of philosophy of science.

Sections 2 through 4 provided a deep dive into the validity associated with three distinct practices: psychometric testing, experimentation in social sciences, and animal modeling for psychiatric disorders. Each of these realms offers its methodological distinctiveness and has its envisioned research implications. While such practices might intertwine, identifying the specific validity associated with a given practice aids in unveiling the diverse interpretations of validity, thereby linking it to its theoretical, historical, and pragmatic contexts. Yet, determining the validity "of" a particular practice does not address the intended purposes steering the validity assessment. By specifying validity "for" a set objective, we can unearth the epistemic foundations and practical objectives that shape our understanding of the term.

In sum, the multifaceted concept of validity is informed by its diverse historical trajectories and mirrors the variety of practices and their objectives found across research traditions. By pinpointing the validity specific to certain practices (validity "of") and intended purposes (validity "for"), we can effectively connect the concept to its comprehensive context, allowing for a clearer understanding and more appropriate applications in various research arenas.

References

Alexandrova, Anna, and Daniel M. Haybron. 2016. "Is Construct Validation Valid?" *Philosophy of Science* 83(5): 1098–1109. https://doi.org/10.1086/687941

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1985. *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.

2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. 1966. *Standards for Educational and Psychological Tests and Manuals*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. 1974. *Standards for Educational and Psychological Tests*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, and National Council on Measurements Used in Education. 1954. *Technical Recommendations for Achievement Tests*. Washington, DC: American Psychological Association.

Ankeny, Rachel A., Sabina Leonelli, Nicole C. Nelson, and Edmund Ramsden. 2014. "Making Organisms Model Human Behavior: Situated Models in North-American Alcohol Research, since 1950." *Science in Context* 27(3): 485–509. https://doi.org/10.1017/S0269889714000155, PubMed: 25233743

Baetu, Tudor M. 2016. "The 'Big Picture': The Problem of Extrapolation in Basic Research." *British Journal for the Philosophy of Science* 67(4): 941–964. https://doi.org/10.1093/bjps/axv018

Belzung, Catherine, and Maël Lemoine. 2011. "Criteria of Validity for Animal Models of Psychiatric Disorders: Focus on Anxiety Disorders and Depression." *Biology of Mood & Anxiety Disorders* 1(1): 9. https://doi.org/10.1186/2045-5380-1-9, PubMed: 22738250

Borsboom, D., A. O. J. Cramer, R. A. Kievit, A. Z. Scholten, and S. Franić. 2009. "The End of Construct Validity." In *The Concept of Validity: Revisions, New Directions, and Applications*, edited by R. W. Lissitz, 135–170. IAP Information Age Publishing.

Borsboom, D., G. J. Mellenbergh, and J. van Heerden. 2004. "The Concept of Validity." *Psychological Review* 111(4): 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061, PubMed: 15482073

Brookshire, David S., Don L. Coursey, and Willam D. Schulze. 1987. "The External Validity of Experimental Economics Techniques: Analysis of Demand Behavior." *Economic Inquiry* 25(2): 239–250. https://doi.org/10.1111/j.1465-7295.1987.tb00737.x

Buckingham, B. R., W. A. McCall, A. S. Otis, H. O. Rugg, M. R. Trabue, and S. A. Courtis. 1921. "Report of the Standardization Committee." *Journal of Educational Research* 4(1): 78–80.

Burt, C. 1924. "Historical Sketch of the Development of Psychological Tests." In *Report of the Consultative Committee on Psychological Tests of*

Educable Capacity and Their Possible Use in the Public System of Education (Hadow Report), edited by Board of Education, 1–61. London: HMSO.
Campbell, Donald T. 1957. "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin* 54(4): 297–312. https://doi.org/10.1037/h0040950, PubMed: 13465924
Campbell, Donald T. 1973. "The Social Scientist as Methodological Servant of the Experimenting Society." *Policy Studies Journal* 2(1): 72–75. https://doi.org/10.1111/j.1541-0072.1973.tb00128.x
Campbell, Donald T. 1986. "Relabeling Internal and External Validity for Applied Social Scientists." *New Directions for Programs Evaluation (Advances in Quasi-Experimental Design and Analysis)* 31: 67–77. https://doi.org/10.1002/ev.1434
Campbell, Donald T. 1991. "Methods for the Experimenting Society." *Evaluation Practice* 12(3): 223–260. https://doi.org/10.1177/109821409101200304
Campbell, Donald T. 1997. "The Experimenting Society." In *The Experimenting Society: Essays in Honor of Donald T. Campbell*, edited by William N. Dunn, 1st ed., 35–68. Abingdon, Oxfordshire: Routledge.
Campbell, Donald T., and Donald W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56(2): 81–105. https://doi.org/10.1037/h0046016, PubMed: 13634291
Campbell, Donald T., and J. C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton, Mifflin and Company.
Carson, John. 1993. "Army Alpha, Army Brass, and the Search for Army Intelligence." *Isis* 84(2): 278–309. https://doi.org/10.1086/356463
Cartwright, Nancy. 2009. "What Is This Thing Called 'Efficacy'?" Pp. 185–206 in *Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511812880.016
Cartwright, Nancy. 2011. "Evidence, External Validity and Explanatory Relevance." Pp. 15–28 in *Philosophy of Science Matters: The Philosophy of Peter Achinstein*. Edited by Gregory J. Morgan. https://doi.org/10.1093/acprof:oso/9780199738625.003.0002
Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press. https://doi.org/10.1093/0195171276.001.0001
Collier, Rex Madison. 1944. "The Effect of Propaganda upon Attitude Following a Critical Examination of the Propaganda Itself." *The Journal of Social Psychology* 20(1): 3–17. https://doi.org/10.1080/00224545.1944.9918827

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

Cronbach, Lee J. 1957. "The Two Disciplines of Scientific Psychology." *American Psychologist* 12(11): 671–684. https://doi.org/10.1037/h0043943

Cronbach, Lee J. 1971. "Test Validation." Pp. 443–507 in *Educational Measurement* 2nd ed. Edited by Robert L. Thorndike. Washington, DC: American Council on Education.

Cronbach, Lee J. 1975. "Beyond the Two Disciplines of Scientific Psychology." *American Psychologist* 30(2): 116–127. https://doi.org/10.1037/h0076829

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52(4): 281–302. https://doi.org/10.1037/h0040957, PubMed: 13245896

Dehue, Trudy. 2001. "Establishing the Experimenting Society: The Historical Origin of Social Experimentation According to the Randomized Controlled Design." *The American Journal of Psychology* 114(2): 283–302. https://doi.org/10.2307/1423518, PubMed: 11430152

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia, eds. 2011. *Cambridge Handbook of Experimental Political Science*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511921452

Feest, Uljana. 2020. "Construct Validity in Psychological Tests—The Case of Implicit Social Cognition." *European Journal for Philosophy of Science* 10(1): 4. https://doi.org/10.1007/s13194-019-0270-8

Feest, Uljana. 2022. "Data Quality, Experimental Artifacts, and the Reactivity of the Psychological Subject Matter." *European Journal for Philosophy of Science* 12(1): 13. https://doi.org/10.1007/s13194-021-00443-9

Fidler, Fiona, and John Wilcox. 2021. "Reproducibility of Scientific Results." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/scientific-reproducibility/

Gage, N. L., ed. 1963. *Handbook of Research on Teaching*. Chicago: Rand McNally.

Gigerenzer, Gerd, Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty, and Lorenz Kruger. 1989. *The Empire of Chance*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511720482

Guala, Francesco. 1999. "The Problem of External Validity (or 'Parallelism') in Experimental Economics." *Social Science Information* 38(4): 555–573. https://doi.org/10.1177/053901899038004003

Guala, Francesco. 2005. *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511614651

Guala, Francesco. 2010. "Extrapolation, Analogy, and Comparative Process Tracing." *Philosophy of Science* 77(5): 1070–1082. https://doi.org/10.1086/656541

Guion, Robert M. 1980. "On Trinitarian Doctrines of Validity." *Professional Psychology* 11(3): 385–398. https://doi.org/10.1037/0735-7028.11.3.385

Henderson, Valerie C., Jonathan Kimmelman, Dean Fergusson, Jeremy M. Grimshaw, and Dan G. Hackam. 2013. "Threats to Validity in the Design and Conduct of Preclinical Efficacy Studies: A Systematic Review of Guidelines for In Vivo Animal Experiments." *PLoS Medicine* 10(7): e1001489. https://doi.org/10.1371/journal.pmed.1001489, PubMed: 23935460

Heukelom, Floris. 2011. "How Validity Travelled to Economic Experimenting." *Journal of Economic Methodology* 18(1): 13–28. https://doi.org/10.1080/1350178X.2011.556435

Hood, S. Brian. 2009. "Validity in Psychological Testing and Scientific Realism." *Theory & Psychology* 19(4): 451–473. https://doi.org/10.1177/0959354309336320

Jiménez-Buedo, María. 2011. "Conceptual Tools for Assessing Experiments: Some Well-Entrenched Confusions Regarding the Internal/External Validity Distinction." *Journal of Economic Methodology* 18(3): 271–282. https://doi.org/10.1080/1350178X.2011.611027

Jiménez-Buedo, María, and Luis Miguel Miller. 2010. "Why a Trade-Off? The Relationship between the External and Internal Validity of Experiments." *THEORIA* 25(3): 301–321. https://doi.org/10.1387/theoria.779

Jiménez-Buedo, María, and Federica Russo. 2021. "Experimental Practices and Objectivity in the Social Sciences: Re-Embedding Construct Validity in the Internal–External Validity Distinction." *Synthese* 199(3–4): 9549–9579. https://doi.org/10.1007/s11229-021-03215-3

Kane, Michael T. 1992. "An Argument-Based Approach to Validity." *Psychological Bulletin* 112(3): 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kazdin, Alan E. 2000. "Campbell, Donald Thomas." In *Encyclopedia of Psychology*. Washington, DC: American Psychological Association; Oxford/New York: Oxford University Press.

Kelley, T. L. 1927. *Interpretation of Educational Measurements*. Oxford: World Book Co.

Landsberger, Henry A. 1957. *Hawthorne Revisited: A Plea for an Open City*. Ithaca, NY: Cornell University.

Mayo, Deborah. 2008. "Some Methodological Issues in Experimental Economics." *Philosophy of Science* 75(5): 633–645. https://doi.org/10.1086/594510

McKinney, William T., and William E. Bunney. 1969. "Animal Model of Depression." *Archives of General Psychiatry* 21(2): 240–248. https://doi.org/10.1001/archpsyc.1969.01740200112015, PubMed: 4980592

Meehl, Paul E. 1978. "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology." *Journal of Consulting and Clinical Psychology* 46(4): 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Messick, Samuel. 1975. "The Standard Problem: Meaning and Values in Measurement and Evaluation." *American Psychologist* 30(10): 955–966. https://doi.org/10.1037/0003-066X.30.10.955

Messick, Samuel. 1980. "Test Validity and the Ethics of Assessment." *American Psychologist* 35(11): 1012–1027. https://doi.org/10.1037/0003-066X.35.11.1012

Michell, Joel. 2013. "Constructs, Inferences, and Mental Measurement." *New Ideas in Psychology* 31(1): 13–21. https://doi.org/10.1016/j.newideapsych.2011.02.004

Mosier, Charles 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, PubMed: 20256558

Mundfrom, Daniel J., Dale G. Shaw, Ann Thomas, Suzanne Young, and Alan D. Moore. 1998. "Introductory Graduate Research Courses: An Examination of the Knowledge Base." In *American Educational Research Association Annual Meeting*.

Nestler, Eric J., and Steven E. Hyman. 2010. "Animal Models of Neuropsychiatric Disorders." *Nature Neuroscience* 13(10): 1161–1169. https://doi.org/10.1038/nn.2647, PubMed: 20877280

Newton, Paul E., and Stuart D. Shaw. 2014. *Validity in Educational and Psychological Assessment*. London: Sage Publications Ltd. https://doi.org/10.4135/9781446288856

Parker, Wendy S. 2020. "Model Evaluation: An Adequacy-for-Purpose View." *Philosophy of Science* 87(3): 457–477. https://doi.org/10.1086/708691

Reiss, Julian. 2019. "Against External Validity." *Synthese* 196(8): 3103–3121. https://doi.org/10.1007/s11229-018-1796-6

Ruch, G. M. 1929. *The Objective or New-Type Examination: An Introduction to Educational Measurement*. Chicago: Scott, Foresman and Co.

Russel, R. W. 1964. "Extrapolation from Animals to Man." In *Animal Behaviour and Drug Action*, edited by H. Steinberg, 410–418. London: Churchill.

Russo, Federica. 2022. *Techno-Scientific Practices: An Informational Approach*. Lanham, MD: Rowman & Littlefield.

Slaney, Kathleen. 2017. *Validating Psychological Constructs*. London: Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-38523-9

Staay, F. Josef van der, Saskia S. Arndt, and Rebecca E. Nordquist. 2009. "Evaluation of Animal Models of Neurobehavioral Disorders." *Behavioral and Brain Functions* 5(1): 1–23. https://doi.org/10.1186/1744-9081-5-11, PubMed: 19243583

Steel, Daniel. 2004. "Social Mechanisms and Causal Inference." *Philosophy of the Social Sciences* 34(1): 55–78. https://doi.org/10.1177/0048393103260775

Steel, Daniel. 2007. *Across the Boundaries*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195331448.001.0001

Steel, Daniel. 2010. "A New Approach to Argument by Analogy: Extrapolation and Chain Graphs." *Philosophy of Science* 77(5): 1058–1069. https://doi.org/10.1086/656543

Stone, Caroline. 2019. "A Defense and Definition of Construct Validity in Psychology." *Philosophy of Science* 86(5): 1250–1261. https://doi.org/10.1086/705567

Sullivan, Jacqueline A. 2009. "The Multiplicity of Experimental Protocols: A Challenge to Reductionist and Non-Reductionist Models of the Unity of Neuroscience." *Synthese* 167(3): 511–539. https://doi.org/10.1007/s11229-008-9389-4

Vernon, P. E. 1963. *Personality Assessment: A Critical Survey*. London: Routledge.

Willner, Paul. 1984. "The Validity of Animal Models of Depression." *Psychopharmacology* 83(1): 1–16. https://doi.org/10.1007/BF00427414, PubMed: 6429692

Willner, Paul. 1986. "Validation Criteria for Animal Models of Human Mental Disorders: Learned Helplessness as a Paradigm Case." *Progress in Neuropsychopharmacology and Biological Psychiatry* 10(6): 677–690. https://doi.org/10.1016/0278-5846(86)90051-5, PubMed: 3809518