Validity Inquiry Through Cronbach's Lens

Professor, Faculty of Lifelong Learning and Career Studies, Hosei University
Tomoyuki Yasuda

1. Introduction

Lee J. Cronbach is best known for a reliability coefficient called Cronbach's a. Cronbach's a indicates the internal consistency of responses underlying a test (Cronbach, 1951). Being able to examine the reliability of test scores based on the data collected at one point in time, Cronbach's reliability coefficients have widely been used among social scientists and psychologists in particular. In addition to gaining popularity for Cronbach's a, his work also influenced the field of psychometrics and program evaluation.

For instance, Cronbach and his colleague, Paul Meehl, provided in-depth explanations as to what construct validity was and introduced the guidelines for establishing the validity. The guidelines became a standard of practice for educational and psychological researchers. Moreover, Cronbach's view of social phenomena that was drastically different from the experimental tradition lead him develop the challenging technical frameworks known as Generalizability Theory.

The purpose of this paper was to report a comprehensive review of Cronbach's work by focusing on validity issues[1]. Given that his way of viewing the social world, or validity inquiry more specifically, changed throughout his entire academic careers, this paper discussed how he changed his views overtime. This paper was thus organized by 1) the perspectives of construct validity and nomological network (e.g., Cronbach & Meehl, 1955), 2) rejections of nomological network and experimental tradition (e.g., Cronbach, 1975; Meehl, 1967), 3) introduction of the alternative framework in social inquiry (e.g., Cronbach, 1982), and finally 4) nature of validity inquiry (Cronbach, 1988; 1986; 1980).

2. Construct Validity and Nomological Network

Technical recommendation by American Psychological Association (APA), American Educational Research Association (AERA), & National Council on Measurement in Education (NCME) in 1954 set forth the four distinct types of validity in assessing the quality of test scores (APA, AERA, & NCME, 1954). These four types were called concurrent, predictive, content, and construct validity.

The first two types of validity can be obtained by contrasting test scores to certain criteria and, hence, are called criterion-related validity. Major differences between concurrent and predictive validity include whether the criterion concerns what happens after a test is administered (i.e., predictive) or what happens when test is administered simultaneously (i.e., concurrent). Moreover, evidence of content validity can be sought by showing adequacy of a sample of items from the universe of items.

Unlike the other three types of validity evidence (i.e., concurrent, predictive, and content validity), lack of consensus existed as to what construct validity was and how to obtain evidence for it. No formal standard was introduced in the technical recommendation; rather, a set of methodologies, including examination of known-group differences, factor structures, and internal consistencies, were adopted by most of the social scientists by convention (Cronbach & Meehl, 1955).

As a result, Cronbach and Meehl proposed a principle of construct validation under what is called the nomological network. The nomological network refers to the interlocking system of laws which constitute a theory, and they argued that the nomological network should logically be connected to a theory that guides every aspect of research processes (Cronbach & Meehl, 1955).

Central to the nomological network were the constructs and the links that connect these constructs. Cronbach and Meehl (1955) believed that to know what something was was to form the lawful network of relationship within which it occurs. A necessary condition of a construct can be established if it is observed within a nomological network. Elaboration occurs if researchers learn more about the network. Modification occurs if they change the network to fit into the observations, accepting the fact that "there may be alternative constructs or ways of organizing the net which for the time being are equally defensible" (Cronbach & Meehl, 1955, p. 290).

In other words, a sufficient condition can never be achieved in studying a nomological network, just like a theory can never be proven. Rather, to find evidence in a nomological network indicates that the network is the best possible explanations of the lawful relationships constituting a theory that has yet to be disproved. Although the law that governs the network may be statistical or deterministic, as Cronbach and Meehl (1955) argue, construct validation requires the accumulation of relevant evidence, and thus never ending.

3. Rejection of Nomological Network and Experimental Tradition

Regardless of these seemingly promising ideas about the logic underlying construct validation, Cronbach and, especially, Meehl argued against the nomological network at a later point in time. Their major criticism (against their own ideas!) concerned the ways the network was studied in social science in general and psychology in particular. Coming from the perspective of philosophy of science, Meehl

(1967) insisted that it was impossible to establish the nomological network in "soft" science because the test of the network was stochastic in nature.

Meehl (1967) argued that stochastic characteristics inherent to the nomological network created problems when testing the links between the networks. To highlight this point, he insisted on keeping a clear distinction between the substantive theory underlying the networks and the statistical hypothesis which was derived from the substantive theory (Meehl, 1967). It was the test of statistical hypothesis, or null hypothesis significance testing more specifically, that was most often used for judging the quality of substantive theory. This was the very fact that Meehl disagreed with the adoption of the nomological network to the field of social science.

He believed that law-like relationships that form the nomological network are "correlations, tendencies, statistical clusterings, increments of probabilities, and altered stochastic dispositions" (Meehl, 1978, p. 813). Thus, the nomological network in social science was "at best an extension of meaning and at worst a misleading corruption of logician's terminology" (p. 813). More crudely, he criticized null hypothesis testing such that:

> "Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in

the history of psychology" (p. 817).

Meehl bases his argument against the use of the null hypothesis on the fact that the point-null hypothesis is always false in social science. It is always false because of the universal agreement that everything is correlated with everything else to begin with. It is thus nonsensical to test a research hypothesis against the null hypothesis. What is worse, according to Meehl, is the common practice that takes place among researchers for the test of a research hypothesis (H_1: $\mu_e > \mu_c$) against the directional null hypothesis (i.e., H_0: $\mu_e \leq \mu_c$).

Given that the point-null hypothesis is always false, if researchers arbitrarily assign one of the two directional hypotheses (i.e., $\mu_e > \mu_c$ or $\mu_e < \mu_c$), then either one of these hypotheses is correct. Thus, a research hypothesis is correct half of the time without having any logical relationship between theoretical expectations and experimental outcomes.

Consequently, Meehl (1967) argued that 'the value $p = \frac{1}{2}$ is a lower bound on the success-frequency of experimental "tests," assuming our experimental design had perfect power" (p. 111). It was because of this uncertainty in testing socio-psychological or educational theories that he rejected construct validation using the nomological network.

Furthermore, Cronbach (1975) expressed his dissatisfactions for the practice of the experimental tradition established by Campbell and Stanley (1963) and Cook and Campbell (1979) on the ground that they failed to look at the ontological complexity existing in the social world.

Coming from the perspective of "correlational tradition," which is different in scope from that of "experimental tradition" (e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979), Cronbach (1957; 1975) proposed the research agenda that takes Aptitude Treatment Interaction (ATI) into consideration.

By aptitude he meant any characteristic of a person affecting his or her response to the treatment. Various characteristics or dimensions of aptitude, together with different dimensions of treatment, could produce interactional effect in outcome. Citing Domino's (1968) study as an example, Cronbach explained ATI that existed between students' learning styles and teaching orientations on achievement. Two different types of students' learning styles included were Ai (achievement via Independence) and Ac (Achievement via Conformance). Students' who possessed high Ai believed that they do good work when they could set tasks for themselves. On the other hand, those who indicated high Ac thought they do well in meeting requirements others set for them.

Matching the teaching orientations, or treatment, that emphasized students' independence (IND) or conformity (CONF) with students' style of learning, or aptitude (i.e., Ai and Ac), it was found that higher levels of achievement were observed for groups matched by high Ai with IND and by high Ac with CONF than those in which high Ai matched with CONF and high Ac matched with IND, evidencing ATI.

What can be recognized in ATI was the importance of "stratification" of a sample in examining outcomes. As discussed earlier, aptitude could have been any characteristics of a person that interact with treatment, and thus the source of stratification should have been more than one.

To address this point, Cronbach cited the Mischel's (1973) argument that in order to predict one's voluntary delay of gratification, researchers may need to know 1) how old a subject is; 2) his or her sex; 3) the experimenter's sex; 4) the particular objects for which the subject is waiting; 5) the consequences of not waiting; 6) the models to whom the subject was just exposed; 7) the subject's immediately prior experience, and so forth. Cronbach argues that these seven variables, or source of stratification, can add up to 120 interactions, which of course is beyond comprehension in practice.

4. Alternative Framework in Social Inquiry

The ATI in social science research, along with the criticism against the experimental tradition, led Cronbach to seek other paths – one of which included the formulation of an alternative framework adopted to the field of program evaluation (e.g., Shadish, Cook, & Leviton, 1991).

Using the term "social inquiry," instead of "social experimentation", he insisted on the importance of looking at the ontological complexity embedded in evaluating a program. In particular, he believed that human behavior could not should not be reduced to a series of parsimonious causal (or relational) laws, because "[W]hen he [or she] says that such-and-such a relation is

true, 'other things being equal,' [she or] he is speaking from the experience of having made a lot of things equal" (Cronbach, 1975, p. 121) – things (i.e., aptitude) that might be of particular importance to the program evaluators.

(1) The UTOS framework

In order to focus attentions on complex elements (e.g., unit of analysis, treatment, and generalizability) responsible for designing program evaluation, Cronbach employed a set of symbols called "UTOS," representing unit (U), treatment (T), observation (O), and settings (S). More specifically, "U" or unit represents the population of persons about which some conclusions are sought; "T" or treatment denotes the plan for the program and its installation; "O" or observation represents the plan for collecting some types of data; and "S" or settings stand for situations in which the study is made.

In reality, instances on which a study (e.g., data collection) was conducted were represented as lower- and upper-case symbols "utoS," where "uto" represents research participants (u), intervention or research treatment (t), and observation (t) obtained out of treatment (i.e., measurement). Notice that Cronbach made settings (S) as upper-case "S," for a setting was unique to each study (situations, time, conditions, etc.).

Furthermore, Cronbach designated "*UTOS" (pronounced "star UTOS") as a situation or class about which a conclusion was made. Specifically, *U represents a set of units not as fully representative of "U," *T represents a set of treatment that

is different from treatment (intervention) that was implemented in a study, *O characterizes outcomes not measured directly in the original study, and *S corresponds to the settings when and where the findings are to be generalized.

Cronbach's (1982) argument originated from the limited reach in traditional internal validity (i.e., sin qua non) as discussed in Campbell and Stanley (1963) and Cook and Campbell (1979). Instead, Cronbach believed that external validity was of primary interest and thus should have been the focus of program evaluation. In making his case, Cronbach explained that the causal statements for which internal validity were to be claimed were extremely restricted because causal conclusions were conditional and thus limited the applications of such conclusions. A social phenomenon did not reside within a stable, isolated system, and hence causal language was loose at best.

(2) Validity under the UTOS framework

Cronbach (1982) viewed internal validity as a statement about UTOS made on the basis of study conducted on utoS. "The inference is internal to UTOS in the sense that utoS is a subset of observations from that universe" (Cronbach, 1982, p. 106). Thus, he viewed internal validity as reproducibility, such that the utoS was for the UTOS.

He continued by saying that internal validity, in Campbell's term, concerned only about "t" (cause) to "o" (effect), making "internal validity a property of trivial, past-tense, and local statements" (Cronbach, 1982, p. 137).

Treatment (T) is embedded in and hence cannot be separated from U, O, and S. We need to speak of causation by paying attention to this fact. Cronbach argued that the homogeneous sample of not only "u" but also "toS" should be sought for the purpose of showing the evidence of causality. Thus, the internal validity from Cronbach's lens involved with whether or not the reported facts on utoS were dependable.

Cronbach (1982) also viewed external validity differently from Campbell and his colleague. In fact, Campbell's external validity required that the research findings could be generalized to the population, hence uto(S) to UTO(S), but in fact this was internal validity for Cronbach (Shadish et al., 1991).

Instead of generalizability, Cronbach (1982) recognized external validity as "extrapolation," reflecting his point of view that this validity signified a deliberate projection to a situation outside the range where information was gathered. Thus, statistical summaries on the sample (utoS), as well as estimates for UTOS, were usually not an adequate base for inference about *UTOS (i.e., external validity evidence).

Questions as to why local "t" worked well or poorly need to be understood in reference to *UTOS in order to maximize external validity. Cronbach believed that the alternative sources of social information, including such qualitative knowledge as history, culture, and folktale, should be incorporated in addressing the differences between utoS and *UTOS. Thus, the evidence of external validity was less systematic and stylized than statistical inference. Specifically, the range of extrapolation needs to be shortened in order to improve the credibility of extrapolation – that is, a part or all of the original utoS should be close to *UTOS. Cronbach argued that careful documentation especially for "S" was considered necessary, given that setting was unique to each study.

5. Validity Inquiry

Having examined various points of view that Cronbach and colleagues subscribed the question remains, how validity inquiry should be conducted.

The comprehensive analyses of validity issues as discussed in psychology and related fields were beyond the scope of this paper[2]. Yet, Cronbach (1980) maintained that validity inquiry (e.g., test validation) calls for integration of many types of evidence; that is, the varieties of investigation are not alternatives and any one of which would be adequate.

Indeed, S. Messick (e.g., Messick, 1995; 1989) later on identified the significance of integrating various evidences when inquiring validity, such that "[T]he unified concept of validity integrates consideration of content, criteria, and consequences into a construct framework for testing rational hypotheses about theoretically relevant relationships" (Messick, 1989, p. 64). It was the construct validity on which this unifying view of validity was based. Accordingly, validity was defined as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and

actions on the basis of test scores or other modes of assessment" (Messick, 1995, p. 741; Messick, 1989)[3].

Cronbach (1988; 1986; 1980) also identified the importance of not only the technical aspect of validity inquiry but also the social aspect of it. Concurrent to Messick's point of view, Cronbach believed that effective communication of evidence to the relevant audience was central to the improvement of social aspects of validity inquiry (i.e., credibility in social inquiry)[4]. Among all, utility was thought of as essential in social inquiry. There were a total of five perspectives as significant for engaging in social inquiry – function, political, operationalist, economic, and explanatory perspectives (Cronbach, 1988).

In explaining functional perspective, first, Cronbach argued "worth" rather than "truth" should be the focus in social inquiry. He also focused attentions on the political perspective. There were issues such as fairness and bias being of primary concern. The operationist perspective included the content and construct on which validity evidence were sought.

From utility point of view, moreover, he argued about the importance of the economic perspective. In particular, what validity evidence was sought and how it was to be used (e.g., cost and benefit) could have become of central interest to researchers, practitioners, and administrators. Finally, the bridge between theoretical and practical point of views was considered critical under the explanatory perspective.

As has been discussed throughout this paper, validity inquiry was to be conducted by taking into consideration not only the technical manipulation of research methods but also the conceptual and rhetorical arguments about the whole spectrum of issues. These issues were situated wide-ranged from psychological to societal in practice. Thus, validity inquiry, in and of itself, required various theories and methods of investigation.

Endnotes

1 This paper was written based on the author's class presentation titled "Evolving and Rhetorical Nature of Validation: A perspective of Lee J. Cronbach" given in the EDPSY 555 class (Validity by Dr. Hoi K. Suen) at the Pennsylvania State University Educational Psychology Ph. D. Program in 2002.

2 Non-traditional yet substantially important approach to validity can be found in an argument-based approach to validity, where observation, generalization, extrapolation, theory-based inferences, decisions, and technical interferences were to be sought in establishing the interpretive argument (Kane, 1992). Early development of the field of program evaluation also identified the importance of validity argument (e.g., House, 1980).

3 It was thus important for researchers to engage in validity inquiry by acknowledging the fact that "the appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that unifying force is empirically grounded construct interpretation" (Messick, 1989, p. 64).

4 Although Cronbach's perspective is strikingly different from that of experimental tradition (e.g., D. T. Campbell's experimental design) as has been discussed throughout this paper, it was Campbell

who also viewed that the knowledge construction in social science should be via the disputatious community of truth seekers – a group of scientist which is "open, mutually reinforcing but critical commentary on the procedures and assumptions underlying any knowledge claim" (Shadish et al., 1991, p. 120).

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological bulletin, 51*, 1-38.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologists, 57*, 671-684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116-126.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New directions for testing and measurement, 5*, 99-108.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., Walker, D. F., Weiner, S. S. (1980). *Toward reform of program evaluation*. San Francisco: Jossey-Bass

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Cronbach, L. J. (1984). *Essentials of Psychological Testing (4th Ed)*. New York: Harper & Row Publishers.

Cronbach, L. J. (1986). Social inquiry by and for earthlings. *In D. Fiske & R. Schweder (Eds.). Metatheory in Social Science* (pp. 83-109). Chicago: University of Chicago.

Cronbach, L. J. (1988). Five perspectives on the validity argument. *In H. Wainer & H. I. Braun (Eds.) Test Validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Domino, G. (1968). Differential prediction of academic achievement in conforming and independent settings. *Journal of Educational Psychology, 59*, 256-260.

House, E. R. (1980). *Evaluating with Validity*. Beverly Hills, CA: Sage.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Meehl, P. E. (1967). Theory-testing in psychological and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Meehl, P. E. (1986). What social scientists don't understand. *In D. Fiske & R. Schweder (Eds.). Metatheory in Social Science* (pp. 315-338). Chicago: University of Chicago.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80*, 252-283.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of Program Evaluation: Theory and practice.* Newbury Park, CA: Sage Publications.

Messick, S.: 1989, 'Validity', in R. L. Linn (ed.), *Educational Measurement*, 3rd ed. (American Council on Education, New York), pp. 13–103.