When the ancient Greek philosopher Bion of Borysthenes (c. 325 – c. 250 BCE) articulated his famous ranine apothegm, he likely did not conduct a validity test to see if the frogs were really dead. To the tableau of three fractious youths pelting innocent frogs, he was a casual observer. Yet, until frogs fly, his interpretative observation may not have valorized reality. For as the empirical skepticist Nietzsche perspectivistically (perspectivism is a theory that perception is influenced by the preceptor) generalized, “there are no facts, only interpretation” (2003, p. 139). Data science may offer dispositive solutions, but – on the other hand, “it is not possible for the scientific panopticon to guarantee rational decision, which is not surprising since there are no guarantees in science” (Messick, 1998, p. 37). Causally, empirical research and quantitative research share the premise that certain “regularities” obtain and can be detected and reproduced recursively (Jackson, 2010, Chapter 10). Comparably inscrutable, the late renowned British statistician George E. P. Box reputedly ribbited, “All models are wrong, but some are useful” (Crane, 2018, p. 8), whereupon Harry Crane, Associate Professor and Co-Director of the Graduate Program in Statistics and Biostatistics at Rutgers University and Fellow at the London Mathematical Institute, croaks: “‘making sense’ is the first step towards ‘being useful’” (ibid.). Likewise, Blumer – though not a Calaveras Countian – leaps to the conclusion, “methods are mere instruments designed to identify and analyze the obdurate character of the empirical world, and as such their value exists in their suitability in enabling this task to be done” (1986, p. 27). Ultimately, scholars – symbolic interactionists, like Blumer, as well as situational social world mappers, ecological catalogers, generalizability theorists [such as Cronbach’s α (Alpha)-ers], multi-stage samplers, structural probabilists, local-to-global elementalists, and like ilk (passim; Crane, 2018; Clarke, 2017; Creswell, 2014; Marsden, 2010; and, Williams, D. E., 2010) – concur that the validity of any survey instrument predicates upon its ability to assess that utility relative to what it purports to measure (Weiner, 2010, p. 1827). While presenting a general definition of validity as the “degree to which an instrument measures the construct under investigation,” Bohrstedt also suggests that a “measure has as many validities has as there are uses to which it is put,” i. e. its instrumental utility (Marsden, 2010, pp. 273-4). Implicit to that utility are the associative values of reliability and validity, which may be defined, respectively, as:

*the degree of consistency with what instances are assigned to the same category by different observers or the same observers on different occasions* (Silverman, 2005) and validity taken from Silverman (2005) who defined it as ‘I mean truth.’

(Williams, D. E., 2010, p. 121-2; see also Fink, 2008, p. 43ff).

Reliability and validity, “the two crucial and universally-accepted criteria in assessing and evaluating the quality of any research” (Tuward, 2017, n. p.), inherently possess a correlative co-efficiency since “reducing or eliminating [threats to validity]…depends upon quality information, strategies for improving reliability can also improve validity” (Coghlan,
Of course, since its introduction in 1951, Lee Cronbach’s α (Alpha) has been the consensus statistical item-total reliability process to “determine the similarity [read: internal consistency] in participants’ answers across all the items in a scale” (Allen, 2017, p. 1415; see also: Cronbach, 1951, pp. 297-334), with scores above .70 achieving high reliability. Cronbach’s famous formula is:

\[
\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}}
\]

where N is equal to the number of items, \(\bar{c}\) is the average inter-item covariance among the items and \(\bar{v}\) equals the average variance (https://stats.idre.ucla.edu/spss/faq/what-does-cronbachs-alpha-mean/). Yet, recent literature depreciates Cronbach's alpha, e.g. problems coincident to measurement of two items over a period of time, such as test-retest reliability or parallel-forms reliability; and, as one expert observed in a private email, Cronbach’s alpha “conflates the number of items with how well they correlate. With N items, the size of the covariance matrix is N*N, whereas the list of variances (comprising the diagonal of the matrix) is N. It’s not hard to see that as N increases, N*N/N will also increase under a lot of conditions – not necessarily the ones you want.” Another formulaic option is the Spearman-Brown Prophesy, which “can predict what the reliability of a scale will be if one adds or deletes items” (Allen, ibid., p. 1417):

\[
\rho'_{xx'} = \frac{n \rho_{xx'}}{1 + (n - 1) \rho_{xx'}}
\]

Apropos such “reliability,” a PhD Coordinator of Institutional Research coincidentally posted the following more grass-roots-based validity observation to the Association Assessment of Learning in Higher Education’s (AALHE) archive:

*I should, however note, that low reliability does not necessarily mean low "VALIDITY," which is the question of interest. Reliability is usually substituted for Validity when evaluating instruments simply because it is far easier to obtain. In fact, the Mental Measurement Yearbook made the point that 95% of all measures lack more than minimal (content) validity. Low reliability precludes extremely high validity, but reliability is nothing more than a measure of the consistency with which scores are generated (theoretically on the same subject assessed multiple times). Validity relates to how closely the scores cluster around a theoretical target TRUE SCORE. Scores may cluster very tightly around the WRONG target and receive extremely high reliability [sic] indices yet be totally INVALID. On the other hand, scores may spread around the desired target and exhibit comparatively low reliability, yet be reasonably VALID particularly when averaged (the average will be closer to the TRUE if they vary in two or more dimensions about the TRUE. Our extensive research on Raters/Observers suggests that three is the optimum number one should use, although given the usual tight monetary situations, two can suffice with only a moderate loss of consistency in evaluation. However, two is an absolute minimum. [link to source]"
In her step-by-step guide to conducting surveys, Fink discusses **qualitative measurement** and **design validity**: the former constituent encompasses the characteristics of the instrument while the latter focuses on its content. Intuitively, development of a survey instrument posits a rational need, a *cui bono?* (“for whose good”) – as one researcher expressed it. Overridingly, the notion of “whose good” constitutes the organizational good, which – perhaps apathetically unobvious to their majority – encompasses the randomized participants in the survey. As the old saying goes, “When you want to drain the marsh, you don't consult the frogs,” since the frogs will manifest a self-serving cognitive bias: ultimately, “the value in value added depends on the ecology” (Braun, article title).

As a *sine qua non*, probabilistic pre-condition, randomization fortifies a survey’s scope condition “since each element in the population is given a known nonzero probability of being selected into the sample” (Marsden, 2010, p. 7).

Validity may be informed by salient internal and external factors as outlined below:

**Table 1: Threats to Internal Validity**

<table>
<thead>
<tr>
<th>Threat to Internal Validity</th>
<th>Definition and Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection of participants</td>
<td>Biases resulting from selection of two groups of respondents that are not equivalent to one another in age, motivation, education, etc. For instance, one respondent group is primarily male and the other is primarily female. This may happen if the groups are not randomly selected or the randomization process does not work.</td>
</tr>
<tr>
<td>History</td>
<td>Unanticipated events that occur while the survey is in progress. For instance, satisfaction with a school-based program to encourage healthier eating may be affected by a healthy-eating campaign on a popular children’s TV show.</td>
</tr>
</tbody>
</table>
| Maturation                  | Processes that occur within participants (e.g. physical and emotional growth) that inevitably occur as a function of time. Children in a 3-
A year school-based physical education program mature physically. For example

| Testing                      | Effect of taking one survey upon the responses and outcomes of a subsequent one. After a 3-week program, for example, respondents are surveyed. They recall their answers on the first survey, and this influences their responses to the second one |
| Instrumentation              | An effect due to changes in the survey instrument or to changes in scores or administrators. For instance, Surveyor I makes slight changes between the questions asked of the respondents over three administrations. Surveyor A administers the questionnaire and gives directions the first two time the survey is administered, but Surveyor B is called in to administer it the third time. |
| Statistical regression       | An effect operating where respondents are selected on the basis of extreme scores and regress or go back toward the mean (e.g., average score) of that variable. Only people at great risk are included in the survey, for example, and some of them inevitably “regress” to the mean or average “score.” Regression to the mean is a statistical artifact (i.e., due to some factor or factors outside of the study). |
| Attrition                    | The differential loss of respondents from one or more groups on a nonrandom basis. For example, busier people don’t complete the survey after the first administration. |

**Table 2: Threats to External Validity**

<table>
<thead>
<tr>
<th>Threat to External Validity</th>
<th>Definition</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interaction effects of selection biases and the experimental treatment</td>
<td>This threat occurs when an intervention or program and the participants are a unique mixture, one that may not be found elsewhere. The threat is most apparent when groups are not randomly constituted. Suppose a large school volunteers to participate in an experimental program to improve the quality of its staff’s leisure time activities. The characteristics of the school’s administration (some of which – like leadership and priorities –</td>
<td>Suppose a large company volunteers to participate in an experimental program to improve the quality of it employees’ leisure time activities. The characteristics of the company (some of which – like leadership and priorities – are related to the fact that it volunteered for the experiment) may interact with the program so that the two together are unique; the particular blend of company</td>
</tr>
</tbody>
</table>
are related to the fact that it volunteered for the experiment) may interact with the program so that the two together are unique. The particular blend of school and program can limit the applicability of the findings.

| Reactive effects of testing | Bias that occurs when a survey is given before an intervention or program resulting in an effect that will not generalize. Two groups of students participate in a research study. Group 1 is given a survey before watching a film, but Group 2 just watches the film. Group 1 performs more positively on the survey that is given to both groups after the film the first survey sensitizes them to the film’s content. | Sometimes known as the Hawthorne effect, this threat is caused when people behave uncharacteristically because they are aware that their circumstances are different. (They are being observed by cameras in the classroom, for instance, or they have been “chosen” for an experiment.) |
| Reactive effects of experimental arrangements or Hawthorne effect | An effect that occurs when respondents know that they are participating in an experiment. |  |
| Multiple program interference | A bias that results when respondents are in other complementary activities or programs that interact. Participants in an experimental mathematics program are also taking physics class. Both teach differential calculus exactly the same way. |  |

construct as “theoretical ideas developed to explain and to organize some aspects of existing knowledge….It is a dimension understood or inferred from its network of interrelationships” (APA, 1974, p. 29). To this clarification, Bohrstedt adds APA’s amplification:

The investigation begins by formulating hypotheses about the characteristics of those who have high scores on the [measure] in contrast to those who have low scores. Taken together, such hypotheses forma at least a tentative theory about the nature of the construct the [measure] is believe to be measuring.

Such hypotheses of theoretical formulations lead to certain predictions about how people...will behave...in certain defined situations. If the investigator’s theory is correct, most of [the] predications should be confirmed” (ibid. p. 30).

If Bion were living now, he could conduct a battery of such tests in order to determine valid results. Among the various versions of validity testing, the following are more or less canonical to the psychometric discipline:

- **construct**: assesses the extent to which an instrument captures a specific hypothetical construct, generally through measurement of its severally substratified “domains of meaning;”
- **content**: assesses the extent to which the instrument adequately samples the specific intended domain of content;
- **criterion (aka predictive or concurrent)**: assesses validity based by comparing scores to another previously validated metric;
- **face**: simplest assessment which measuring the instrument’s tacit meaning.

(Kline, 2015, p. 6ff; Little, 2013, pp. 343-5; Litwin, 1995, p. 35ff; Marsden, 2010, p. 372ff; and, Weiner, 2010, pp. 1827-8)

Apposite the substratification of the construct domains of meaning, Bohrstedt observes, “The more refined the strata and substrata, the easier it is to construct items late, and the more complete the coverage of meanings associated with the construct” (ibid, p. 375).

These epistemological distinctions may only be significant to psychometrians, psychoeducators, or other specialists. The “Introduction” to one specialist resource, the Mental Measurement Yearbook (MMY), testifies that this longstanding (1938-) series “provides factual information, critical reviews, and comprehensive bibliographic references on the construction, use, and validity of all tests published in English” (Buros, 1965, n. p.). But, notably, even these specialists have “for years…complained that publishers of standardized achievement tests overemphasize content validity and neglect construct validity” (Buros, 1989, p. 131).

In his book on library administration and organization, Williams posits that “another method used to ensure trustworthiness an rigor is through the use of [data] triangulation...“the use of more than one data source,” such as the library triangulating its already-validated, third-party Ruffalo® Noel-Levitz’ data results against its proprietary survey data results for the identical two Association of College and Research Library-normed service and collections outcomes, viz.:
Guidelines for University Library Services to Undergraduate Students

Services

Effective, high quality undergraduate library services successfully support the undergraduate programs of the institution.

* * *

Resources and collections

The library should provide varied, authoritative and up-to-date resources that support its mission and the needs of undergraduate users.

http://www.ala.org/acrl/standards/ulsundergraduate

Another overarching research-based validation process involves the Angoff method, which “relies on subject-matter experts (SMEs) who examine the content of each test question (item) and then predict how many minimally-qualified candidates would answer the item correctly. The average of the judges’ predictions for a test question becomes its predicted difficulty” (ALTA Language Services, Oct 6, 2008, n. p.).

Bhaskaran provides an insightful example of the directionality that a validity question might take:

* If you want to know about a respondent’s social networking use, it wouldn’t make sense to ask how many social networks he belongs to. Belonging to a social network is not a valid indicator of social network usage. Many people join numerous social networks and never participate in the conversation. By the same token, many join only one social network and engage with the community throughout the day. Validity in this case would be to ask directly about the amount of time the respondent spends using social networks.

(2010, p. 91).

Upon observing Leonidas, Dionysius and Apollonius cold-bloodedly pelting the knot of frogs – ostensibly to their annihilation, Bion could have examined the frogs to determine their morbidity. Wishing to close the assessment loop (see: Banta, 2011) but perhaps fearing warts from actually touching the frogs, Bion may have alternatively given the three rock throwers a survey assessing the extent to which they thought the frogs were dead. However, as we have seen, surveying brings its own radically allied set of validity issues. Bion’s survey questions would have been subject to each of the three individual’s parametric psychodynamics, which renowned Danish mathematician, statistician, and psychometrician Georg Rasch notes as “the extent to which the item elicits the latent trait and the position on the subject on this trait” (Kline, 2015, p. xi).

To test instrument validity, Kline suggests correlation with survey(s) criteria using the following process:
(1) Set out precisely hypotheses concerning the observed and latent variables with which the test should correlate (concurrent or criterion-based validity).
(2) Set out precisely hypotheses concerning the variables with which the test should not correlate.
(3) Specify groups, which should score high and groups, which should score low on the test.
(4) Hypothesize the place of the test in factor space. This is a similar hypothesis to those of (1) and (2) above.

(ibtid. 2015, p. 157)

Similarly, Fink postulates the following expedients for testing these types of survey validity:

- concurrent validity “by comparing it [the survey] against a known and accepted measure;”
- content validity “by proving that its items or questions accurately present the characteristics or attitudes they are intended to measure;”
- construct validity “by demonstrating that they measure a construct such as hostility or satisfaction.”

Methodologically, all survey risks should be assessed prior to implementation of the instrument to respondents in order to ensure its validity. Selection bias can be a major causal pollutant to any survey question development. In conclusion, the following illustration sums up the systematized stepwise protocols designed to enhance validity:

**Figure 2: How Test Validity Works**
Although the following statement may be subject to validity testing, truly: “No frogs were harmed in the production of this essay.”

“Brekekekex koax koax,”¹ which might be Greek for “this test is valid.”

NEXT QUARTER’S ARTICLE will focus on Implementation Fidelity.

References


