when is a bathroom scale unlike a political poll? when it's reliable.

by Ruth M. Corbin

In judging market research, courts of law consider at least three criteria: reliability, validity and pertinence. The definitions of these terms will be familiar to most readers. Reliability refers to the reproducibility of results and, for survey research, is significantly governed by the quality of the sampling process. Validity refers to the ability of a measurement instrument to measure accurately the intended object of analysis. Variations in interpretation are sometimes given through terms such as "external validity", "predictive validity", "construct validity", "face validity", or "convergent validity." The third criterion, pertinence, refers to the relevance of the outcome to the issue in dispute, and is usually best left to the lawyers to battle out.

Social scientists are more likely to face questions of reliability and validity than other measurement experts. Physical characteristics, studied by physiological scientists, are relatively easy to calibrate without so much concern for complex error. If you need to measure the height of your growing child, for example, you need only put a yardstick against the wall, and record the number of centimeters. Social scientists are usually interested in more complex intangible matters such as attitudes and dispositions that have no rigid yardsticks. The yardsticks they have invented, such as aptitude tests, attitude scales, and political polls, are vulnerable to measurement errors that can be challenging to calibrate.

The legendary Howard Cosell, in the late 70s, was said to have complained into a microphone as Philadelphia Phillie shortstop Larry Bowa entered the batter's box: "His batting average is only 261, but this kid is a 300 hitter."[1] Cosell was expressing a perceived discrepancy between the true ability of a batter and the aptitude test (the batting score) that was supposed to capture it. Researchers call it measurement error. The word "error" in their context is not intended to have negative connotation. It need not imply that a mistake has been made. Measurement of intangibles, and even of many tangibles, will always entail some inaccuracy or error. The duty of scientists is to estimate the magnitude of the error, and, of course, to strive to control it.

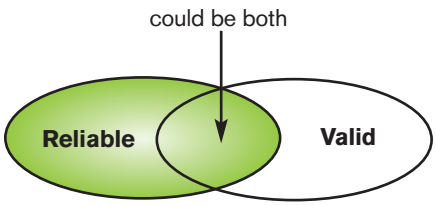
The theory of measurement error arose in the field of psychology, to account for the imperfect predictability of aptitude test scores. British psychologist Charles Spearman is credited with first adapting the statistical concept of reliability in 1904 to psychological test scores, in his article entitled "The Proof and Measurement of Association between Two Things" in the *American Journal of Psychology.* His work was quickly advanced by Edward Thorndike, Spearman and Thorndike being well recognized as architects of reliability theory in psychological testing. They mostly addressed reliability on a per person basis. The type of question they tackled, for example, was whether a test of a person's aptitudes, measured on different occasions, would produce similar results. If it did, it was reliable; if it didn't, it wasn't. In the mid 20th century, when surveys emerged, sociologists stretched the concept of reliability to refer to repeated measures not on individuals, but on populations.

But evolution of social science has brought with it an interesting dilemma. There appears to be disagreement among different social science professionals as to whether reliability and validity can be

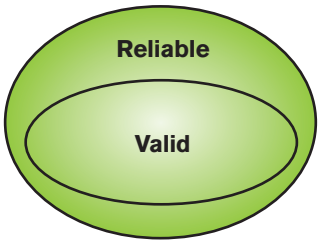
independent criteria, or totally interdependent. There are two points of view:

• Proponents of the "independence" view say that reliability and validity can be evaluated separately. A measurement or study can be valid but not reliable, or reliable but not valid. A typical example given to illustrate this concept is a bathroom scale. If you keep your bathroom scale at a starting point of 5 pounds (perhaps a perverse motivator for losing weight), one would say that your bathroom scale is reliable in always telling you the same weight, but invalid because it never tells you your correct weight. Similarly, say proponents of the independence view, a study can be valid but not reliable. A prototype example given to illustrate this concept is an exit survey of 100 voters from a particular voting booth between the hours of 3 p.m. and 5 p.m. In the hands of a skilled interviewer, such a survey might be valid – that is, the voters might be willing to tell the interviewer the truth of how they voted. However, as a survey of Canadian voters, the survey would not be reliable, because the 100 daytime voters in a single riding are unlikely to be representative of the voting population at large. In other words, it would be a valid indicator of how those citizens had voted, but not a reliable indicator of the outcome of the election. The book *Trial by Survey*[2] enumerates the inconsistent treatment in case law of surveys which may be valid but not reliable. The treatment in courts of law has ranged from acceptance to rejection or to some middle ground where the survey is given partial weight in the judge's deliberations.

The view that a study can be one or the other (or both or neither) might be illustrated as follows:



• Proponents of the "interdependence" view say that, while a study can be reliable but not valid, if it is unreliable then it must be invalid. Put another way, they say, a study cannot be both valid and unreliable. Validity can only be accompanied by reliability. Examples given to illustrate this point are typically about tests on individual people: if an aptitude test never gives the same score for an individual, then obviously it can't be a valid indicator of that person's ability. The view that validity implies reliability (but not vice versa) may be illustrated as follows:



Is this disagreement among professionals just a language nuance, or a difference of fundamental opinion? It must be just a language nuance, because, as illustrated above, each party to the debate is capable of giving persuasive examples that complement its own definitions.

So, if it is just a language or definitional matter, why call for consistency? Why not say "to each his/her own definition" and be done with it? The resolution pursued in this article is written for those who care about precision of language as terms of art. It is also written for students who want to go on believing that published textbooks have correct information. Finally, it is written for judges who need to assess whether conflicting experts are debating something of substance or merely of definition.

The inconsistent treatment of the concept of reliability seems to have arisen when different groups of social scientists each evolved the concept to their own priorities and measurement tools, specifically those falling into the following four categories:

1. The first category includes individual tests on individual people, such as aptitude tests, eye examinations, or fitness tests. The reliability or "reproducibility" of the result for a given individual might be imperfect because the result may differ according to different random circumstances or noise factors, that would influence a person's score on any given day. Reliability is measured by a Test-Retest Correlation, that is by correlating scores obtained on different occasions. Assuming that the time that had elapsed between the first and second tests was irrelevant to the thing being measured, and that appropriate scientific controls had been exercised, the test result could not be valid if it proved unreliable.

2. A second category covers tests which incorporate different versions such as the LSAT or MCAT tests, the aptitude exams for law schools and university medical programs. Reliability in such cases is measured by an Alternative-Forms Reliability Correlation, that is by calculating correlations among different occasions of test administration to the same people (if that were feasible), or by calculating correlations among test results of "like individuals." Reliability is typically a reflection on the test format assuming comparable and representative test items, and quality controls over administration. In this category, like the last one, if the test results are unreliable across different versions, then they cannot be collectively valid.

3. A third category covers measurement scales, where a bank of items is used to compile an overall rating of a target construct, as might be done in measuring student satisfaction with a particular professor. In designing a scale of this type, each item is thought to measure some aspect of the overall construct, and to contribute to a summed score. The items should therefore exhibit internal consistency, in what they say about the construct in question. A Split-Half Reliability test may be used to measure the internal consistency or reliability of the scale. The items are randomly divided into two groups, and results for a large number of individuals are corre-

lated. Reliability of the scale items would be confirmed by a high degree of correlation between the two groups of items, that is the internal consistency between them. In the absence of split-half reliability, the scale could still be valid, if it turned out that the items were actually measuring a multi-dimensional construct, with different items measuring different dimensions.

4. A fourth category involves sample surveys of populations. A predominant question of interest to survey scientists is whether the results of a particular survey would be repeated (within a margin of error) if tried on a different sample of people. The focus is on the reliability of the sampling process, rather than the questionnaire. Indeed, in some cases the reliability of the questionnaire is not even meaningful to them. A survey, they say, is just a snapshot in time. An individual's attitude may indeed change over time, and that is to be expected. A voting intention at the early part of an election, expressed to a pollster, may not be a person's same voting intention on the eve of the election. It would be inappropriate to "blame" the questionnaire for its apparent unreliability. In summary, when it comes to surveys, reliability is better applied to the sampling process, because reliability of the questionnaire instrument is frequently less meaningful. In this category, the study results may be valid (the questionnaire measured what it was supposed to measure), but could prove unreliable (if the sampling is not properly done.)

Table 1

Type of study or measure	Type of reliability analysis available	Relationship between reliability and validity	What assumptions underlie this conclusion?
Results of a single test on individuals.	Test-retest correlation or matched group correlation.	If not shown to be reliable then individual test cannot be valid.	Assume random noise and extraneous factors are controlled. If matched groups are used, assume the matching is as close as reasonable to being perfect.
Results of a test with different versions.	Alternative Forms reliability correlation	If not shown to be reliable, then test format cannot be valid.	Assume different versions are comparable, each with representative items capturing the construct of interest. Assume random noise and extraneous factors are controlled. Assume comparable samples are tested for each version.
Results of a measurement scale of a construct, according to which scores on different items are weighted and summed.	Split-half reliability test	If not shown to be reliable, then measurement may still be valid.	One starts by assuming that the construct being measured is uni-dimensional, to warrant summing across different items of a single scale. The truth of this assumption cannot be known until after the results are captured. Thus apparent lack of "reliability" as measured by the split-half reliability test does not automatically rule out validity.
Results of a typical survey of attitudes, opinions or intentions.	Reliability measured by the margin of error attached to the sampling process.	Can be reliable and not valid or valid and not reliable or both or neither.	Use of standard margin of error statistics assumes a random sampling process. If this assumption does not apply, the results may still be valid for the population actually captured.

CONCLUSIONS

1. Reliability of a measurement result refers to its reproducibility or consistency. Systematic sources of error may undermine validity, but do not adversely affect reliability, because they impact the measurement in a constant way, and need not lead to inconsistency.[3]

2. When it comes to surveys, reliability and validity can be separately determined, because they can depend on separate components of the overall survey process. They may not actually be independent of each other in every study, but the "general case" is one in which they can be evaluated separately. Thus, the view that an invalid study can never be reliable is too restrictive. Counterexamples given earlier confirm that such a restrictive view is untenable.

3. Table 1 summarizes the discussion regarding when reliability and validity are inter-dependent and when they are not.

4. This analysis should help to clarify, for students, the bewildering inconsistencies they may find in different textbooks.

5. The analysis may also be helpful to research practitioners in constructing check-lists for quality controls (validity and reliability to be separately considered on such a list!) and in choosing the best analytic option for assessing reliability.

6. Finally, the analysis is intended to be helpful to judges and other triers of fact in assessing critiques of expert evidence. An expert opposing a survey has two separate burdens of proof on the issues of reliability and validity. An expert cannot (as they are sometimes wont to do) dismiss a study as being "totally unreliable and therefore invalid."

7. The significance to a court is that a valid study – even if not representative of the entire pertinent population – can sometimes be probative of a materially important segment of the population, and be therefore deserving of some weight. How much weight to give it depends on its complementarity with other evidence, the unique facts of the case, and ultimately, of course, judicial discretion.

[1] Retold in Kaplan, R. M. and Saccuzzo, D. P. *Psychological Testing. Principles, Applications and Issues.* Monterey: Brooks/Cole Publishing Company, 1982, p. 85

[2] Corbin, R., Gill, A.K. and Jolliffe, R.S., *Trial by Survey.* Toronto: Carswell, 2000

[3] For a straightforward discussion on this point, see, e.g. Malhotra, N. *Marketing Research. An Applied Orientation.* Upper Saddle River, NJ: Pearson Prentice Hall, 2002

Ruth M. Corbin is CEO of Corbin Partners Inc. and Adjunct Professor at Osgoode Hall Law School. She can be reached at rcorbin@corbinpartners.com or (416) 413-7600.