Let’s Not Overlook Content Validity

Manus Rungtusanatham, School of Business, University of Wisconsin–Madison

Scientific knowledge in the Operations Management (OM) discipline, as we know, has traditionally been derived from outcomes of deductive, modeling-based research using either optimization or simulation methodologies. However, such optimization-based or simulation-based modeling research approaches are no longer the only modes of knowledge generation within OM. Today, more and more OM scholars (myself included) have employed or are employing empirical research designs in order to address core issues and problems in OM. Furthermore, an increasing number of OM-based journals have switched or are switching from exclusively publishing modeling-based OM research to also publish empirical OM research (e.g., Journal of Operations Management—see Ebert, 1990).

If we were to compare contemporary empirical OM research with research conducted in the early 1980s, we would, no doubt, admire the remarkable progress that has been made during this two-decade span. Our progress is evidenced not only by the quantity, but also the quality and sophistication of the research endeavors that have been completed. Inarguably, this progress reflects our ever-increasing appreciation for and knowledge about empirical research design, execution, and methodologies.

In this regard, we have benefited first from examining and critiquing the research of pioneer OM scholars who took the first plunge, many of whom are senior members of the Decision Sciences Institute. Their leadership and willingness to enhance the OM research paradigm at a critical junction in the development of OM as a science has allowed empirical OM research to flourish. I have often wondered what paths I, myself, might have professionally and personally traversed had empirical OM research been strongly discouraged during my years in the doctoral program at the University of Minnesota.

At the same time, we have also benefited from formally and informally educating ourselves as to the strengths and pitfalls of conducting high quality empirical research. At the University of Minnesota, for example, doctoral students interested in empirical research are encouraged to supplement their operations research “toolkit” with courses in social sciences research methodologies. Scholars within, as well as from outside, of the OM discipline have further contributed to our education by writing about issues involving the proper design and execution of empirical OM research. The article by Flynn et al. (1990) comes to mind immediately. More recently, Dröge (1996, 1997) has enlightened us to the issues of measurement quality.

As we have matured in our understanding, we have paid increasing attention to the parallel issues of measurement quality and of quantitative assessments of reliability and construct validity in the conduct of empirical OM research. In the case of multiple multi-item measurement scales administered in survey questionnaires, we would compute Cronbach’s $\alpha$ (Cronbach, 1951) and employ factor analysis, either exploratory or confirmatory, to demonstrate that these measurement scales have some degree of reliability and construct validity. It has now become the norm to report such assessments of reliability and construct validity for measurement instruments—whether they be questionnaires, interview protocols, observer checklists, etc.—in papers published in OM journals or presented at various OM conferences.

While reliability and construct validity are important issues of measurement quality, there is, however, another equally critical issue that I believe we should be concerned with, namely the issue of content validity in the operationalization of
What is Content Validity?

Content validity is, therefore, one type of validity. More specifically, the content validity of a measurement instrument for a theoretical construct reflects the degree to which the measurement instrument spans the domain of the construct’s theoretical definition; it is the extent to which a measurement instrument captures the different facets of a construct. In theory, a measurement instrument designed to measure a specific construct has content validity if the items in the measurement instrument constitute a randomly chosen subset of the universe of items that represent the construct’s entire domain. As such, the purpose of assessing an instrument’s content validity can be stated in the form of the following question: “Is the substance . . . of this [measurement instrument] representative of the content or universe of content of the [construct] being measured?” (Kerlinger, 1973, p. 458).

To be able to answer this question, we presume that it is convenient and possible to specify, and to randomly sample from, the universe of items reflecting the construct’s domain. This presumption is, of course, rarely, if at all, satisfied in practice. Consequently, assessments of content validity have typically relied on “appeals to reason regarding the adequacy with which important content has been sampled and on the adequacy with which the content has been cast in the form of [measurement] items” (Nunnally, 1967, p. 82). Our acceptance of such appeals, in lieu of additional evidence of content validity, essentially grants permission to the researcher(s) creating the measurement instrument to define the domain of the construct being measured. Of course, to the extent that we are able to support our appeals by citing appropriate literature lends greater credibility to our reasoning and conclusion of content validity.

What Is Face Validity and How Is It Related To Content Validity?

When a measurement instrument has been created to operationalize a particular theoretical OM construct, and assuming that we are willing to accept the measurement instrument as representing the construct’s theoretical domain, then the assessment of content validity can be satisfied by evaluating the face validity of the measurement instrument. Nunnally (1967, p. 99) defined the face validity of a measurement instrument to be judgments about a measurement instrument after it has been constructed to operationalize a theoretical construct. These judgments focus on the degree to which items in a measurement instrument appear, on their face value, to measure the single construct that they intend to measure.

Regarding the relationship between content validity and face validity, there are fundamentally two camps of thoughts. Some scholars see face validity as different and separate from content validity (e.g., DeVellis, 1991; Kerlinger, 1973). Others (e.g., Carmines and Zeller, 1979; Nunnally, 1967) consider content validity and face validity to be two sides of the same coin and view the assessment of a measurement instrument’s face validity to be an indirect approach to the assessment of content validity. It is the latter perspective to which I subscribe and which allows for a quantitative assessment of content validity.

How Can Content/Face Validity Be Assessed Quantitatively?

In my own research, I have found at least two different approaches for assessing face validity: the Content Validity Ratio (Lawshe, 1975) and Cohen’s (1970) k.
Content Validity Ratio (CVR)

In this approach, a panel of subject-matter-experts (SMEs) is asked to indicate whether or not a measurement item in a set of other measurement items is “essential” to the operationalization of a theoretical construct. The SME input is then used to compute the CVR for each ith candidate item in a measurement instrument (CVR) as follows:

\[ CVR_i = \frac{n_e - N}{2}, \]

where

- \( CVR_i \) = CVR value for the ith measurement item,
- \( n_e \) = number of SMEs indicating a measurement item is “essential,” and
- \( N \) = Total number of SMEs in the panel.

We can infer from the CVR equation that it takes on values between -1.00 and +1.00, where a CVR = 0.00 means that 50% of the SMEs in the panel of size N believe that a measurement item is “essential.” A CVR > 0.00 would, therefore, indicate that more than half of the SMEs believe that a particular measurement item is “essential” and, thereby, face valid. Lawshe (1975, p. 568) has further established minimum CVR’s for different panel sizes based on a one-tailed test at the \( \alpha = 0.05 \) significance level. For example, if 25 SMEs constitute the panel, then measurement items for a specific construct, whose CVR values are less than 0.37, would be deemed as not “essential” and would be deleted from subsequent consideration.

An example of using this approach can be found in Collard’s (1992) development of a measurement instrument for the 14 Points in the Deming Management Method.

Cohen’s (1960) \( \kappa \)

In this approach, J SMEs are asked to independently sort N independent measurement items into an exhaustive set of C a priori defined and mutually exclusive measurement scales for different constructs. Based on the classifications by the J SMEs, we can, then, assess the degree of inter-expert agreement as to the placement of these measurement items into their measurement scales by computing and evaluating Cohen’s \( \kappa \) as follows:

\[ \kappa = F_a - F_c \]
\[ N - F_c \]

where

- \( F_a \) = number of measurement items classified into the same categories by all J judges, summed over all categories \( i \) for \( i = [1, \ldots, C] \). So,

\[ F_a = \sum_{i=1}^{C} F_{i(a)} \]

and

- \( F_c \) = number of measurement items for which agreement, as to their classifications, among all J judges is expected by chance, again summed over all categories \( i \) for \( i = [1, \ldots, C] \). So,

\[ F_c = \sum_{i=1}^{C} F_{i(c)} \]

with

\[ F_{i(c)} = N \cdot \left( \prod_{j=1}^{J} \frac{F_{ij}}{N} \right) \quad \forall i \]

and

\[ F_{ij} \] = number of measurement items classified into ith category by the jth judge.

The obtained values of Cohen’s \( \kappa \) will range from +1.00 to -1.00, where Cohen’s \( \kappa > 0.00 \) means that the observed agreement among the judges is beyond chance agreement. Cohen’s \( \kappa = +1.00 \), therefore, signals perfect inter-judge agreement. Cohen (1960, p. 42) pointed out that the case of Cohen’s \( \kappa < 0.00 \) is “likely to be of no further practical interest . . . ,” since the observed agreement is less than expected by chance.

Cohen (1960) also proposed an approximation to the standard error of Cohen’s \( \kappa \), \( \sigma_\kappa \), that can be computed in the following manner:

\[ \sigma_\kappa = \sqrt{\frac{F_c(N - F_a)}{N(N - F_c)^2}} \cdot \sqrt{\frac{1}{N} - \frac{F_a}{N}}. \]

With a large \( N \) (the number of measurement items to be sorted), the sampling distribution of Cohen’s \( \kappa \) will approximate normality by the Central Limit Theorem. Therefore, confidence intervals for can be constructed for Cohen’s \( \kappa \). A test for the significance of Cohen’s \( \kappa \) can also be conducted, where the null hypothesis, \( H_o \), specifies Cohen’s \( \kappa = 0.00 \). Failure to reject the null hypothesis signifies that the computed Cohen’s \( \kappa \) arose in sampling from a population of measurement items for which inter-judge agreement is a result of chance only. In order to conduct the test for \( H_o: \) Cohen’s \( \kappa = 0.00 \), the test statistic, \( z_\kappa \), is calculated as follows:

\[ z_\kappa = \frac{Cohen’s \kappa}{\sigma_\kappa} \]

where

\[ \sigma_\kappa = \sqrt{\frac{F_c(N - F_a)}{N(N - F_c)^2}} \cdot \sqrt{\frac{1}{N} - \frac{F_a}{N}} \]

denotes \( \sigma_\kappa \) when \( F_a \) is constrained to be zero.

The \( p \) value of \( z_\kappa \) on the corresponding normal curve can then be determined in the usual manner.

For an example of using Cohen’s \( \kappa \) in an assessment of content validity, please see Rungtusanatham, Anderson, and Dooley (forthcoming).

Conclusions

The importance of establishing measurement quality as an integral part of the conduct of empirical research has been stressed.
As I pointed out earlier, we have heeded such advice and are paying increasing attention to measurement quality, particularly regarding issues of reliability and construct validity, in the conduct of empirical OM research. However, because a measurement instrument with no content validity will not operationalize a theoretical construct of interest, I sincerely hope that we will begin assessments of measurement quality by quantifying the content validity of measurement instruments used in our empirical research. We need to go beyond invocations of literature support in offering rigorous evidence of content validity. Literature support, in my opinion, is a necessary but insufficient condition for concluding content validity.

References


Dr. Shawnee Vickery Department of Management, College of Business, 239 Eppley Center, Michigan State University, East Lansing, MI 48824, 517-353-6381, fax: 517-432-1112 email: 22645skv@msu.edu

Tony Nolan & Colin Innes, the DIGEST is a collection of stand alone modules that cover over 300 different aspects of decision making, tools and diagrams/models.

In early 1998 this digest will be available free to any DSI who would like a copy on disk. If you would like further details, or would like to find out more about the Decision Intelligence Group at the University of Technology, please contact:

Tony Nolan School of Management Faculty of Business University of Technology, Sydney PO Box 123 Broadway NSW 2007 Australia email: t.nolan@uts.edu.au