■ SHAWNEE VICKERY, Feature Editor, Eli Broad Graduate School of Management, Michigan State University

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 today. 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 α (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

Manus Rungtusanatham

This coming Fall 1998, Professor Manus (Johnny) Rungtusanatham will be joining the faculty of the Department of Management at Arizona State University. Before that, he has spent three years as an assistant professor of business in the School of Business at the University of Wisconsin. He graduated in 1995 with a Ph.D. in business administration from the University of Minnesota–Twin Cities, where he majored in operations management and completed his dissertation in the area of quality management. His research has been published in such journals as the Academy of Management Review, Decision Sciences, Journal of Quality Management, Journal of Operations Management, International Journal of Quality & Reliability Management, Engineering Management Journal, *and others. Professor Rungtusanatham currently sits on the editorial review boards of* Journal of Operations Management, Journal of Quality Management, *and* Quality Management Journal *and regularly reviews for* Decision Sciences *and* Academy of Management Journal. *He has presented papers and has served as a discussant for numerous papers presented at annual and international meetings of the Decision Sciences Institute. He has been involved, both as a student and as a panelist, in the Doctoral Student Consortium at past annual meetings of the Decision Sciences Institute. He recently completed serving a two-year term on the Doctoral Affairs Committee and is currently serving a two-year term on the Publications Committee of the Decision Sciences Institute. Dr. Rungtusanatham is also a founding member of the MBA Roundtable, a nonprofit organization dedicated to advancing the quality of MBA programs.*
(manus@bus.wisc.edu)

theoretical OM constructs. An initial survey of empirical OM research being undertaken by two colleagues and myself reveals that we generally downplay the issue of *content validity*. We would often assume *content validity* to be present as an outcome of the process in which we have constructed our measurement instruments. Or, in instances where we do pay attention to *content validity*, we would appeal to reason or invoke (and justify) literature support to serve as evidence of *content validity*. Rarely do we attempt to quantify the degree of *content validity* in our operationalizations of theoretical OM constructs, perhaps because of the misconception that the subjectivity of *content validity* cannot be quantified!

I sincerely believe that assessments of *reliability* and *construct validity* are virtually meaningless unless we have ensured, with some degree of confidence, the *content validity* of measurement instruments that are being proposed for theoretical OM constructs. The reason is that no degree of *reliability* and *construct validity* can compensate for lack of *content validity*. In this forum, I, therefore, hope to bring greater clarity not only to the importance of *content validity* but also to its formal quantitative assessment as part of the empirical research process. To do so, I pose and answer three related questions as follows:

1. What is *content validity*?
2. What is *face validity* and how is it related to *content validity*?
3. How can *content/face validity* be assessed quantitatively as part of the empirical research process?

I trust that answers to these three questions will supplement and extend earlier articles by Flynn et al. (1990) and Dröge (1996, 1997) and further our knowledge about measurement quality.

What Is V*alidity?*

Permit me to begin by reviewing what *validity* is and how *content validity* relates to *validity*. Generally speaking, a measurement instrument's *validity* is "epitomized by the question: Are we measuring what we think we are measuring?" (Kerlinger, 1973, p. 457) and refers to the extent to which an instrument actually measures what it alleges to measure (Carmines and Zeller, 1979, pp. 12, 17). *Validity* focuses attention on the "extent of matching, congruence, or 'goodness of fit' between an operational definition and the [construct] it is purported to measure" (Singleton et al., 1993, p. 115). The assessment of a measurement instrument's *validity*, in this sense, corresponds to an evaluation of the accuracy and adequacy of the measurement instrument as an operational definition for a particular construct (DeVellis, 1991, p. 43). *Validity*, however, "cannot be assessed directly" (Singleton et al., 1993, p. 121) and can only be "inferred from the manner in which [a measurement instrument] was constructed [i.e., *content validity*], its ability to predict specific events [i.e., *criterion-related validity*], or its relationships to measures of other constructs [i.e., *construct validity*]" (DeVellis, 1991, p. 43).

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) κ.

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 *i*th candidate item in a measurement instrument (CVR_i) as follows:

$$CVR_i = \frac{n_e - \frac{N}{2}}{\frac{N}{2}},$$

where

CVR_i = CVR value for the *i*th 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) κ

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 κ as follows:

$$\kappa = \frac{F_a - F_c}{N - F_c},$$

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

$F_{i(a)}$ = number of measurement items classified into the same category by all *J* judges;

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

$$F_{i(c)} = N \bullet \left(\prod_{j=1}^{J} \frac{F_{ij}}{N} \right) \qquad \forall i$$

and

F_{ij} = number of measurement items classified into *i*th category by the *j*th judge.

The obtained values of Cohen's κ 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 κ, σ_κ, that can be computed in the following manner:

$$\sigma_\kappa = \sqrt{\frac{F_a(N - F_a)}{N(N - F_c)^2}} = \frac{\sqrt{F_a\left(1 - \frac{F_a}{N}\right)}}{N - F_c}.$$

With a large *N* (the number of measurement items to be sorted), the sampling distribution of Cohen's κ will approximate normality by the Central Limit Theorem. Therefore, confidence intervals for can be constructed for Cohen's κ. A test for the significance of Cohen's κ 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 κ 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_κ, is calculated as follows:

$$z_k = \frac{\text{Cohen's } \kappa}{\sigma_{\kappa_0}}$$

where

$$\sigma_{\kappa_0} = \sqrt{\frac{F_c}{N(N - F_c)}}$$

denotes σ_κ when F_a is constrained to be zero.

The *p* value of z_κ on the corresponding normal curve can then be determined in the usual manner.

For an example of using Cohen's κ 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 (see Dröge, 1996, 1997). Carmines and Zeller, for example, argued that only when measurement instruments for theoretical constructs are reliable and valid:

> [Can] analysis of [data collected using these measurement instruments] . . . lead to useful inferences about the relationships among the underlying [constructs]. . . . On the other hand, if the theoretical [constructs] have no empirical referents, then the empirical tenability of the theory must remain unknown. . . . [When measurements are unreliable and / or invalid] . . . analysis of [such data] . . . can lead possibly to incorrect inferences and misleading conclusions concerning the underlying [constructs] . . . . (1979, pp. 10-11)

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

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Newbury Park, CA: Sage Publications.

Collard, E. F. N. (1992). *The impact of Deming quality management on interdepartmental cooperation*. Unpublished doctoral dissertation. University of Minnesota.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(4), 297-334.

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Applied Social Research Methods Series, Volume 26. Newbury Park, CA: Sage Publications.

Dröge, C. (1996). How valid are measurements? *Decision Line, 27*(5), 10-12.

Dröge, C. (1997). Assessments of validity. *Decision Line, 28*(1), 10-12.

Ebert, R. J. (1990). Announcement on empirical / field-based methodologies in JOM. *Journal of Operations Management, 90*(1), 135-137.

Flynn, B. B., Sakakibara, S., Schroeder, R. G., Bates, K. B., & Flynn, E. J. (1990). Empirical research methods in operations management. *Journal of Operations Management, 9*(2), 250-284.

Kerlinger, F. N. (1973). *Foundations of behavioral research* (2nd ed.). New York: Holt, Rhinehart, and Winston.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563-575.

McCullough, P. M. (1988). *Development and validation of an instrument to measure adherence to Deming's philosophy of quality improvement*. Unpublished doctoral dissertation. University of Tennessee at Knoxville.

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.

Rungtuanatham, M., Anderson, J. C., & Dooley, K. J. (Forthcoming). Towards measuring the 'SPC implementation / practice' construct: Some evidence of measurement quality. *International Journal of Quality and Reliability Management*.

Singleton, R. A., Jr., Straits, B. C., & Straits, M. M. (1993). *Approaches to social research* (2nd ed.). New York: Oxford University Press. ■

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

Nolan's DIGEST

(Decision Intelligence Generic Educational Self Tutorial) is a group or individual, self-directed flexible learning tutorial package using mix and match modules. By authors 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.