Research Validation: Challenges and Opportunities in the Construction Domain

Gunnar Lucko, Ph.D., A.M.ASCE[1]; and Eddy M. Rojas, Ph.D., A.M.ASCE[2]

Abstract: Validation of the research methodology and its results is a fundamental element of the process of scholarly endeavor. Approaches used for construction engineering and management research have included experiments, surveys and observational studies, modeling and simulation, theory building, case studies, and various subtypes thereof. Some studies use more than one approach. A particular contribution of this paper is that it reviews different types of validation using examples of studies, analyzes the specific challenges that were found to be significant, and presents how they were successfully overcome in each case. Another contribution is that it describes new opportunities for research validation that are emerging at the horizon as well as ongoing collaborative efforts to enhance the access of construction researchers to validation tools. This paper increases the awareness of the paramount role that validation techniques play in scholarly work by providing readers with recommendations to properly validate their own research efforts.

Introduction

Construction engineering and management research examines real-world means and methods in an effort to improve the effectiveness and efficiency of the construction industry. It is essential for any scientific inquiry that researchers ensure the quality of their work in every step of its methodology, including data collection, analysis, and interpretation of results, through appropriate validation techniques. The results and the process by which they were derived need to be accepted by the academic and the professional communities so that the new knowledge becomes another stepping stone in the advancement of the state-of-the-art and filters down to daily practice, ultimately contributing to the welfare of society. It is therefore also important that researchers collaborate with industry practitioners to establish credibility.

The *scientific method* is a structured approach to generate new knowledge through critical inquiry. While the methodologies of scientific research vary significantly depending on the particular subject and domain, its singular goal always remains the same, to discover truth. However, absolute scientific truth is impossible to attain with human means, because scientific endeavor is by definition never-ending and iterative. The large body of knowledge in philosophy that explores the concept of truth underlines its fundamental nature, but is outside of the scope of this paper. As such, the purpose of *validation* is to ensure that each phase of the chosen research methodology rigorously adheres to the highest standards of quality. This level of quality in planning, executing, and evaluating research is measured as validity.

Law (2007) stressed that validation always depends on the specific purpose of the research. El-Diraby and O'Connor (2004) asserted that "no single definition of the ingredients or subsets of the concept of validity" exists. Numerous writers discuss validity, yet each study uses a slightly different definition of the concept (e.g., Roschke 1994; Kamat and Martínez 2003). Sargent (1991) elaborated further that "currently no algorithms or procedures [are] available to identify specific validation techniques, statistical tests, etc., to use in the validation process." Therefore, validation is a challenge to all researchers, but especially so to those working in interdisciplinary fields such as construction engineering and management.

In light of the absence of definite criteria for how to accomplish validation, this paper first provides an overview of the various types of validity. The focus is not on particular research methods, because their inherent diversity renders a one-on-one matching less meaningful, and because several different validation techniques may be used in the same study. Rather, this paper examines various accepted validation techniques and illustrates how the challenge of establishing validity has been successfully resolved in construction research with numerous examples. It also introduces new validation opportunities emerging at the horizon as well as collaborative efforts by various organizations to increase the accessibility of validation resources for construction researchers.

[1]Assistant Professor of Civil Engineering and Director of Construction Engineering and Management Program, Dept. of Civil Engineering, The Catholic Univ. of America, 620 Michigan Ave. NE, Washington, D.C. 20064 (corresponding author). E-mail: lucko@cua.edu

[2]Professor and Executive Director, Pacific Northwest Center for Construction Research and Education, Dept. of Construction Management, Univ. of Washington, 130G Architecture Hall, Box 351610, Seattle, WA 98195. E-mail: er@u.washington.edu

JOURNAL OF CONSTRUCTION ENGINEERING AND MANAGEMENT © ASCE / JANUARY 2010 / **127**

J. Constr. Eng. Manage., 2010, 136(1): 127-135

Validation and Related Concepts

Validation stands in contrast to *verification*. Both terms are common in technical and managerial literature, but are frequently used in an unclear or even interchangeable manner. Verification establishes the technical correctness of a research product, e.g., a simulation model (Sargent 1991), in accordance with its specifications that were established a priori. It is often performed internally by the researcher, who thoroughly checks all components in a manner of proofreading. Errors that should be identified and eliminated during verification can range from trivial and innocuous deviations to essential and destructive divergences. Examples are "bugs" in computer code, transcription errors in survey answers, transposition errors in numerical data, misconnections of schedule activities, and omitted or doubled components. Validation is the more elusive one of the two concepts, as verification is primarily a comparison of measurable performance parameters. Together, they constitute an acceptance by both its developers and users as being credible, which particularly in the area of simulation is called *accreditation* or *certification* (Law 2007).

In general terms, one could state that verification is concerned with "doing things right," while validation is concerned with "doing the right things." The process of validation can be broadly divided into two main areas: establishing *internal* and *external validity*. Internal validity is related to the concept of causality and is preoccupied with the derivability of relations within data (Leedy and Ormrod 2001). External validity is related to the concept of induction and focuses on the generalizability of results for prediction purposes (Leedy and Ormrod 2001). Internal validity can be threatened by many problems, including ill-defined theoretical models that include spurious relationships or correlated explanatory variables, biases in data collection that render comparisons ineffective, and failure to entertain alternative explanations during data analysis. Note that a true causality can only be established under carefully controlled, laboratory-like environment, but not in observational studies, where "only" correlation can be identified. External validity can be threatened by a variety of issues, including lack of statistical rigor in the selection of sample sizes and in collecting actual data, the presence of any special circumstances during the research efforts, and oversimplification of the phenomena under study.

Besides internal and external, other types of validity that are commonly referred to in the literature include *face validity*, *content validity*, *criterion validity*, and *construct validity*. Face validity is a subjective judgment of nonstatistical nature that seeks the opinion of nonresearchers regarding the validity of a particular study (Leedy and Ormrod 2001). For example, a simulation model of an earthmoving operation can be said to have face validity if earthmoving experts looking at the model and its output agree that it represents to a high degree what happens in reality. In a practical field of study such as construction engineering and management, the collaboration with appropriate representatives from the private and public sectors, e.g., industry practitioners, government agencies, and also the public at large, is very important to secure the face validity of the research endeavors.

Content validity is another nonstatistical approach that focuses on determining if the content of a study fairly represents reality. Its primary concern is "the degree to which a measure covers the range of meanings included within the concept" (Babbie 1990). For example, a study of stress levels of construction supervisors is said to have content validity if the tasks, activities, events, and environmental factors that are included can be considered to be representative of those that generally occur in the larger industry.

Criterion validity has been defined as "the extent to which the results of an assessment instrument correlate with another, presumably related measure (the latter is called the *criterion*")" (Babbie 1990). Criterion validity is thus established when the findings of a research study on a particular aspect of the construction industry, e.g., factors that influence craft productivity, agree in general with the outcomes of related studies, which e.g., investigated training programs for skill development or the motivational effects of working conditions, even though their detailed approaches may differ.

Construct validity refers to whether operationalizations of theoretical constructs are appropriate. In other words, construct validity is concerned with ensuring that a research effort is measuring what it is supposed to measure according to its stated objectives (Leedy and Ormrod 2001). For example, a study about the relationship between the motivation and productivity of construction workers is said to have construct validity if, and only if, the instruments used actually measure motivation and productivity. Even though this may seem obvious, there are many "abstract" constructs, especially traits of the human behavior or psyche that have varying definitions without universally accepted metrics or that need to be detailed further to be usable. Motivation, for example, may have to be refined with more objective factors, e.g., need for work, salary and consequences of nonperformance. In such cases, researchers must justify the process by which the instruments and their conceptual contents were developed to ensure construct validity.

Finally, another term that is sometimes confused with validity is *reliability*. It is related to the concepts of consistency and repeatability in the data collection. According to Neuman (2000), it "is necessary for validity and is easier to achieve than validity." As with validity, there are different kinds of reliability, including *interrater reliability, internal reliability, test-retest reliability*, and *parallel forms reliability*. Interrater reliability is concerned with consistency between different measured subjects. For example, a productivity study that relies on several individuals to visit construction sites and collect labor productivity data has interrater reliability if all observers classify the activities of workers in the same manner, i.e., do not impress personal views onto the data, and can thus be considered neutral or "interchangeable" in the analysis.

Internal reliability focuses on the consistency within a measuring instrument, i.e., individual questions thereof, and can be quantified statistically using Cronbach's α (Cronbach 1951). For example, if a questionnaire is created to survey the aforementioned laborers' motivation, it should contain several questions that "cross-check" each other without being apparent to the respondent, i.e., are not just rewordings. Randomizing the order of questions can further remove the respondent's attention from this internal feature. Rich literature on considerations for survey studies can be consulted by construction researchers for designing their measuring instruments (e.g., Babbie 1990; Cresswell 1994; Leedy and Ormrod 2001).

Test-retest reliability refers to the consistency of results when the same measurement tool is repeatedly applied to the same scenario. This repeatability of measurements is not only essential to remove the manner of applying a test as a potential error source, but also if other researchers want to reproduce the documented results. For example, a survey instrument is deemed to have test-retest reliability if the responses are closely correlated when the survey is applied to the same sample persons in two different instances. However, researchers must take care that a possible learning effect by the test subjects, who may remember details of

the previous test, is not mistakenly interpreted as test-retest reliability.

Parallel forms reliability refers to the consistency of different, but related, measurement tools when applied to the same sample. As an illustration, parallel forms reliability can be achieved if two different surveys to assess the level of construction workers' job satisfaction provide highly correlated results.

Construction engineering and management research must be verified, validated, and reliable to achieve the highest level of quality. This paper, however, focuses mainly on the validation process.

Challenges

Given the interdisciplinary nature of construction engineering and management research as well as the inherent difficulties of performing studies in real-life settings, research validation in the construction domain encounters a variety of challenges. This section presents examples of how such typical difficulties have been successfully overcome by construction researchers. In their entirety, these examples provide guidance to the research community for enhancing the quality of construction engineering and management studies by delivering meaningful valid outcomes.

Internal Validity

As mentioned, internal validity is related to causality. Establishing causality is a significant challenge for construction researchers, given that many studies are performed in real-life settings where multiple variables interact with one another, many of them uncontrollable or unpredictable. While laboratory experiments are less common in the construction engineering and management domain than in other areas of inquiry, such as e.g., structural engineering, they represent one of the best strategies to study causality. The tight control of environmental factors that is possible in a laboratory setting can strongly establish internal validity of the research. Researchers should try to perform laboratory studies whenever possible. Ideally, the laboratory study should be augmented with field trials, as performed by Wu (2005) while examining the flowable backfill material for pipelines.

Internal validity can also be achieved by means other than laboratory trials under controlled conditions. For example, researchers can use multiple regression analysis to isolate the effect of individual independent variables within a model (Montgomery et al. 2001). The results of multiple regression analysis, however, must exhibit internal consistency reliability as a necessary condition to be valid. Statistical resampling techniques can be used to ascertain internal consistency reliability without the need to collect new data. Smaller samples are repeatedly taken from an existing data set, provided its sufficient sample size. *Cross-validation* is a well-known technique that can test the stability of a model. The existing data set is randomly split into two portions and the independent and dependent variables of the first one are used to estimate the model coefficients. The fitted model then makes predictions from the independent variables of the second portion, which are compared statistically with its actual dependent variables as shown in Fig. 1. Zayed and Halpin (2005a) split their data set of pile construction productivities into 70% estimation and 30% prediction portions. Lucko et al. (2006) employed a 50% to 50% split after checking Snee's (1977) recommendation of having at least $n \geq 2 \cdot p + 25$ data points to allow splitting, where $n =$ sample size and $p =$ number of estimated parameters in the re-
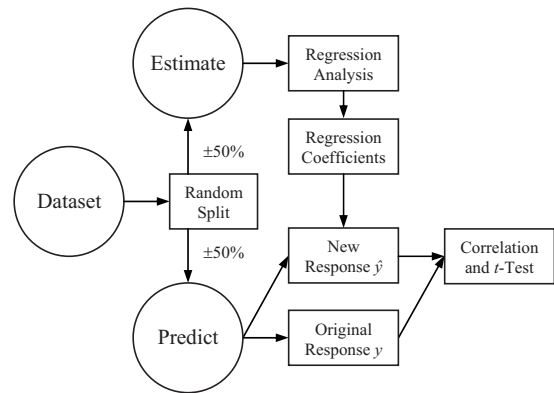

Fig. 1. Cross-validation technique applied by Lucko et al. (2006)

gression model. Menches and Hanna (2006) used "leave-one-out cross validation, [where] one case is left out of the data set, and the remaining $n-1$ cases are used to construct the model" for their performance index for project managers, similar to the resampling technique *jackknifing*. Multiple cross-validations yielded a misclassification rate, which was desired to be less than 10%. Resampling is also employed in *bootstrapping*, named after "pulling oneself up by one's own bootstraps," i.e., with existing data. It assumes that the data set is representative of the population. Numerous new data sets of the same sample size n are created from the existing data set by repeated resampling with replacement. Different from validation, the technique estimates distributions for statistics of interest, e.g., mean or median (Barton and Schruben 2001).

Multiple regression analysis must also satisfy a variety of conditions other than internal consistency reliability to ensure internal validity (Lucko et al. 2006). The internal validity of such studies can be threatened by issues such as misspecification of the model components (excluding important or including unimportant independent variables), multicollinearity (independent variables are highly correlated and thus redundant), micronumerosity (small sample size), heteroskedasticity (nonconstant variance in data points), autocorrelation (repetitive pattern within the data points), nonzero expected disturbances (regression assumption of zero error mean does not hold), and nonlinearity (regression assumption of linear relationship between independent and dependent variables does not hold) (Gujarati 1995). A variety of tests and techniques can be found in the literature to ensure that these problems are not present within a model. Zayed and Halpin (2005b) performed F-tests on their regression models for durations, productivity, and cost of pile construction and t-tests for the null hypothesis that each coefficient β was identical to zero, whose rejections together and individually established the linear relationship of the variables.

Researchers often attempt to define models based on statistical techniques for goodness-of-fit. While these techniques can be very useful in building a functioning model, researchers must still verify that all included variables are related to the dependent variable by some theoretical construct in order to ensure internal validity. In other words, there must be a clear and straightforward reason why a given variable belongs in a model, e.g., ambient temperature and labor productivity, other than just simple correlation with the dependent variable, e.g., the consumption of gasoline and the number of housing starts in residential construction. Otherwise, the variables would be judged as spurious, i.e., seem-

ingly related but in truth both caused by another, hitherto hidden third variable, e.g., the income and debt levels of private households.

External Validity

External validity is related to generalizability of results. In order to generalize from a sample, the sample must be representative of the population. One of the most common means to ensure representativeness is to randomize the sampling procedure. In a survey study on leadership characteristics among construction employees, Ozorovskaja et al. (2007) selected companies at random from listings in the yellow pages. This ensured that all listed companies had an equal chance of being included in their sample. Consequently, external validity was established, i.e., conclusions drawn from the sample could be generalized to that population. In some cases, other techniques are also employed to support generalizability. Stratification divides the population into groups that shall be represented in the sample. Davis and Songer (2003) used "company size, industry sector, and profession, age, and gender" as strata in their mail survey on resistance to technological change to ensure that the proportions of the Unites States construction industry were reflected in their samples. However, conscious sampling alone is not sufficient for external validity. The sample size also determines whether the inherent variability within the population has been captured sufficiently. Depending on whether the study is cross-sectional or longitudinal, the sample size describes the breadth or length of the data set.

Statistical equations provide guidance on necessary minimum sample sizes for research studies, beyond the obvious rule-of-thumb that more samples are always better, especially for very heterogeneous populations (Leedy and Ormrod 2001). Note that this differs from the resolution or "fineness" of the data points, which should be appropriate to the desired level of detail of the research results. Lucko et al. (2006) evaluated the sample size n of their existing data sets of heavy equipment resale values using several recommendations from the literature that included the number of explanatory variables k in the model, e.g., $n \geq 15 \cdot k$ for multiple linear regression, $n \geq 50 + 8 \cdot k$ for multiple correlation, and $n \geq 104 + k$ for testing explanatory variables. These guidelines consider the "level of detail" of the model by counting its variables. The population size also should be considered in determining the sample size a priori. Gay et al. (2006) provided sample size percentages for ranges of population sizes based on earlier research that had assumed a statistical significance level $\alpha = 0.05$ (Krejcie and Morgan 1970)

- The larger the population size, the smaller the percentage of the population required to get a representative sample.
- For smaller populations, say $N = 100$ or fewer, there is little point in sampling; survey the entire population.
- If the population size is around 500 (give or take 100), 50% of the population should be sampled.
- If the population size is around 1,500, 20% should be sampled.
- Beyond a certain point (about $N = 5,000$), the population size is almost irrelevant, and a sample size of 400 will be adequate.

Other writers also give different recommendations depending on whether the data are continuous or discrete (Bartlett et al. 2001). Since sample size is intimately related to various statistical properties, e.g., "the variance, the desired significance level, the power of hypothesis tests . . . and the widths of the confidence and prediction intervals" (Lucko et al. 2006), the writers recommend to confer with a statistician to ensure the external validity of a new research project.

If data are not measured directly by researchers but rather are created by respondents through a measuring instrument, e.g., a survey questionnaire, the response rate becomes especially important. As the ratio of actually received responses divided by randomly selected and requested responses, it directly influences the sample size. Achieving a high response rate is a challenge that construction researchers often face. However, simple techniques can boost the response rate and—by extension—increase their external validity. Davis and Songer (2003) used a combination of techniques, starting with telephone and personal announcements of the survey followed by mailing "a cover letter introducing the research study, an informed consent form, and the questionnaire, along with an envelope in which to place the completed questionnaire to be either picked up later or mailed back . . . Follow-up phone calls and office/site visits were performed as reminders for unreturned surveys," which yielded a response rate of 58.4%. Other techniques can seek to increase the motivation of respondents, e.g., by promising an executive summary of the results (Ling 2002) or a small remuneration, gift, or prize lottery in exchange for completing the survey.

Research is by definition an incremental process and researchers need to be keenly aware of the extent to which their work contributes to the body of knowledge. Describing the scope and limitations of the overall study is as important as revealing its assumptions, possible bias, and possible error sources. External as well as content and face validity directly stem from the research being placed into its proper context while clearly acknowledging what it can accomplish versus what remains for future study. Rojas and Kell (2008) described the limitations of their empirical comparison of project delivery systems in terms of geographic area, industry sector, time frame, convenience sampling from existing government records, and how the sample size related to the Gaussian distribution assumption of the central limit theorem for statistical tests.

The most pertinent and arguably also most difficult form of external and face validation that can be accomplished is the deployment and application of a new model, method, or technology into the ongoing industry practice. It is sometimes called pilot testing or prototyping. Since the construction industry is known to adopt innovations only hesitantly (Davis and Songer 2003), this can present a formidable challenge. Examples of construction research that reported actual implementation include e.g., owner-contractor work structures (Anderson et al. 2004) and change management (Price and Chahal 2006) in industry practice, and in the educational domain the interactive, collaborative, and computer-supported teaching of estimating and scheduling concepts (Jaafari et al. 2001).

Face Validity

Face validity requires the "approval" of nonresearchers regarding the validity of a study. Performing a research study directly under the harsh and somewhat unpredictable conditions of a construction site can generate face validity, provided that such environmental factors in the widest sense are measured and documented as well. Thomas et al. (1999), for example, studied the influence of weather on labor productivity for steel erection projects in the field. Navon and Berkovich (2005) evaluated a new materials management system on a commercial building project, which also allowed a comparison with the existing approach. While a laboratory setting works well for certain studies, e.g., testing new materials, the aforementioned practice-oriented nature of con-

struction research often makes such a real-life setting necessary. Even though this makes creating internal validity challenging, it goes a long way toward securing face validity.

Arguably the strongest way to establish face validity is the involvement of domain experts, also known as subject matter experts, before (a priori), during, after (a posteriori), or throughout the research. Such participation can range from an advisory capacity to active collaboration. Obviously, the earlier and the more such input is sought, the stronger it can contribute to face validity. For example, a hallmark of research projects sponsored by the Construction Industry Institute has been to include approximately a dozen industry representatives in addition to the academic personnel. Gibson et al. (1995) even reported the involvement "of experienced industry professionals representing 16 companies, which ensured expertise for model development" in their study on preproject planning. Especially for the construction industry, which is often viewed as resistant to innovation (Blayse and Manley 2004), it is increasingly essential that through such involvement it takes a more active role in initiating, supporting, and implementing the research that seeks to provide solutions to its current and future needs.

Another technique useful to establish face validity is the performance of interviews. El-Diraby and O'Connor (2004) conducted numerous structured interviews with government agencies and consultants for their bridge plans evaluation model. They deliberately sought a diverse group to attain a wide range of opinions. Considerations for conducting meaningful interviews are similar to those for surveys. Structured questions provide internal reliability. However, interviews allow a richer feedback, as the interviewer can clarify and extend individual items ad hoc in a semi-structured manner. An audio recording or at least detailed notes during or directly after the session is vital. Note that an approval of research with human subjects is typically required.

"Focus groups are carefully planned discussions stimulated within a predefined group environment to obtain perceptions about a defined area of interest in a permissive, nonjudgmental environment" (Yu et al. 2006). A more extensive technique than interviews, they include "interaction and self-disclosure among participants" and thus can lead to a consensus among experts (Yu et al. 2006). However, they need to be carefully moderated to not deviate into tangential conversations. Refining the technique, Yu et al. (2006) created transcriptions of their sessions on critical success factors and performed a content analysis on them. The latter technique is described in detail in the literature of the social sciences, where diverse qualitative data are often handled.

The *Delphi analysis*, named after the ancient Greek oracle, is a validation technique that formalizes the collection of expert feedback. Its characterizing feature is that undergoes at least two iterations. The first phase often collects data through a survey or a group of interviews with experts, which are compiled by the researcher. In the second phase the same experts receive all data and are asked to provide further comments. This "peer review" can also generate a consensus. del Cano and de la Cruz (2002) used a Delphi analysis with interviews for validation in their study of project risk management. Demographic information, such as the age, education, position, and experience with different project types of the interviewed experts as well as their company sizes and industry sectors were reported meticulously to support the validation process.

The *Turing test*, which asks a user to decide whether the anonymous "conversation" partner is a human or a machine, is still uncommon in the literature on construction research but its well known in the area of artificial intelligence. Considered "the

most widely used qualitative method for validating expert systems," Roschke (1994) noted that the "blind" Turing test conceals the identity of the decision-makers and thus reduces bias. It can also establish interrater reliability. For his expert system that selects bridge rails, he asked an expert to review fifteen field cases and for each one to distinguish the computer-generated decision from those of two human experts and concluded that the expert's inability to do so established face validity.

A technique that is specifically applicable for face validation of discrete-event simulation computer models is scrutinizing its trace file (Sargent 1991), also known as event log, to match with the sequence of events as expected by the model developer. This automatically generated record contains times at which any state changes occur. Such instances are activities changing between being idle and active and resources being sent and received. Following the entries in the trace file for at least one complete run and comparing them manually with the execution of the model can reveal small errors in links between model elements, their start and finish conditions, and the consistency of time and resource units as well as large errors in the logic of the entire system. Such manual technique also contributes to the verification of the simulation model.

Beyond checking the bare text of the trace file for validation, it also finds an important use as an input file for simulation postprocessing. Visualization and animation can help users in naturally spotting inconsistencies in the simulation that have eluded the developers (Kamat and Martínez 2003). For users to accept the model as a truthful representation of reality, its behavior must mirror their "mental image" of the real system. Research has been performed on both ends of the spectrum, from simple abstractions as in the animation of icons (Zhang et al. 2002) to almost photorealistic environments of construction operations (Kamat and Martínez 2003). A particular advantage of four-dimensional visualizations, i.e., animations, is that it lets users perform face validation even of complex logic without the need to learn a specific computer interface (Ioannou and Martínez 1996). However, González-Quevedo (1995) also reminds that realism alone does not guarantee a "good" model and that trivialities that may obscure its abstract essence should be avoided. Recent research is turning the previous one-way data flow from simulation to visualization into a dynamic interaction, i.e., users can manipulate the underlying simulation model via the graphical user interface and directly observe the impact of such changes (Rekapalli and Martínez 2007).

Construction research often creates face validity through case studies, meaning the discussion of a self-contained, detailed example application of a new model or method. The issues surrounding the assumptions, scope, limitations, and resolution of the case study are similar to those encountered in modeling. The challenge of striking a balance between realism and abstraction as well as between inclusiveness and manageability becomes a matter of parsimony, applying "Occam's razor," the principle of being "as simple as possible (but as complicated as necessary)" Addressing this issue, Karam et al. (2007) used a case study of a 2-km-long tunnel in Korea to illustrate and validate their computer tool for decision-making for tunnel construction. Rojas and Dossick (2008) presented a comprehensive case study of an integrated construction research and education center that exemplified their experiential learning and interdisciplinary research approaches. Different from their use in research validation, case studies can also be found in the form of scholarly reports on

innovations in construction engineering and management practice, e.g., on physically moving an existing concrete building with hydraulics (Telem et al. 2006).

Content Validity

Content validity relates the content of a study with reality. Collecting data from a population for scientific study is the phase within the research methodology that has a fundamental influence on the quality of the subsequent analysis and interpretation. Since collecting data are time and cost intensive, researchers typically are not able to measure the spectrum of possible observations, i.e., the entire population. Rather, they collect samples in such a way that they are *representative* of the characteristics of the population. If sampling is performed consciously as to which data are collected and how, it can establish significant content validity, e.g., for survey research.

In rare cases it may be possible to aggregate existing data from several redundant sources. Such "triangulation" echoes its nautical meaning of increasing the accuracy by gauging from several angles. It was used by Lucko et al. (2006), who combined data from two independent equipment auction databases to substantiate their content validity. Data received a source label and were merged and sorted. In a two-step mechanism a computer macro then matched entries by sales date, location, and serial number to identify identical events that had been recorded by both sources, filled any gaps with pieces of information from the respective other source, and deleted all doublettes. In a discussion of combining qualitative and quantitative research methodologies, Creswell (1994) uses the term triangulation in the broader meaning of removing "bias inherent in particular data sources, investigator, and method" by combining several of these listed items.

One of the most important ways in which researchers can show the integrity of their research methodology is documenting their entire approach in detail with an open and self-critical mind. Abstracting and modeling real-world processes entails simplifying assumptions, which must be clearly stated so that readers can independently assess the quality of the study, creating content and face validity. Neuman (2000) illustrated the behavior of data points with a bull's eye target, where the scatter (or variability) around the center represents reliability and the drift (or tendency) off the center represents validity. Scatterplots of the raw data show their centrality and any patterns or trends in an approximate form before any statistical analysis is performed. A systematic *bias* would be visible as a recurring deviation from the center. Other meanings of bias may also be encountered, e.g., in a statistical sense as the opposite of random sampling and in seeking statistical estimators to be unbiased. Even the researchers themselves may be biased, e.g., if they unconsciously favor a particular outcome. A related phenomenon is known as publication bias. Conscientious researchers document and seek to remove bias as much as possible. Davis and Songer (2003) acknowledged that "the pilot sample is biased" because companies that were unlisted in the telephone book would have been ignored. They identified another bias toward sampling from managerial positions, as field personnel were less likely to have office addresses, nor answer long surveys. Since participation was voluntary, respondents were generally "more likely to be persons to be inclined to fill out surveys . . . or those that thought that the research sounded interesting" (Davis 2004). A survey was used instead of interviews to "provide a high external validity, low internal validity solution" (Davis 2004).

Criterion Validity

Criterion validity requires correlation between related measurements. A powerful, yet simple technique for criterion validity and reproducibility of research is to compare predictions with those obtained from a different model or method. This can be combined with extending the often less complex nature of such previous model or method to new capabilities, e.g., increased accuracy or wider applicability. An abundance of different types of comparisons is found in construction engineering and management research, e.g., González-Quevedo et al. (1993), who specifically compared the performance of two simulation systems for modeling construction operations in terms of productivity and duration. Lucko (2007) validated the results from a new analytical model for linear schedules with manual calculations. Elazouni and Metwally (2007) used genetic algorithms to automatically optimize schedules under financial constraints and validated the results of their model by performing linear programming on the same data.

Construct Validity

Construct validity ensures that a research effort measures what it is supposed to measure. One way to accomplish construct validity is through the performance of a pilot test. Different from the pilot implementation of research results, the word pilot in this context refers to fine-tuning the instrument before its use in the actual data collection. Ling (2002) established construct validity—and also content validity—in this manner for a study on architects' and engineers' performance. Col Debella and Ries (2006) examined project delivery methods and as the first step of their survey study pilot tested the questionnaire on several respondents to ensure its effectiveness. No numeric measures exist to capture the improvement thus gained, but conceptually the effect is similar to involving domain experts during the course of the research.

Reliability

Reliability requires consistency and repeatability. An interesting approach to establish test-retest reliability of a new approach was used by Kallantzis et al. (2007), who created a random input generator for their linear scheduling research. Ranges of constraints and buffers and a set of possible productivity values were randomized into 25 sample schedules with six sequential activities and six units of work so that "not only is the margin for error reduced, but also the possibility of a coincidental or biased result is minimized and statistic conclusions can be drawn with relative safety" (Kallantzis et al. 2007). In rare cases, it may also be possible to use existing standardized inputs to evaluate the performance of new approaches, e.g., the project scheduling problem library with 2,040 randomly generated resource-constrained networks of four different sizes that was developed by Kolisch and Sprecher (1996). Vanhoucke (2006) used it as a benchmark for examining work continuity in repetitive projects.

A powerful technique to test the internal consistency reliability of models is sensitivity analysis. It seeks to quantify how variability in the values of independent variable impacts the dependent variable (Elmaghraby 2000). In its most basic form, it entails varying each input either randomly or systematically while holding other inputs fixed. Sargent (1991) described a related technique of testing simulation models under extreme conditions, i.e., specifically setting inputs to their possible maxima and minima. A generalized version would be a factorial experimental design (Kelton and Barton 2003). Since inputs may not just vary by

themselves alone, but can possibly create interaction effects, it can be sensible to extend the approach to a scenario analysis, where various combinations of input values are examined in concert. The choice of the specific values depends on the number of inputs, their range and order of magnitude, continuous or discrete nature, and the desired resolution of the output variability. This technique was adapted by Rojas and Aramvareekul (2003), who traced apparent variability in macroeconomic measures of construction labor productivity to actual changes in the weights between industry sectors, each with an "inherent" level of productivity, rather than to a perceived decline of employee output. In an interesting variation of sensitivity analysis of simulation models, Sargent (1991) lists degeneracy as a validation technique, which also entails "removing portions of the model" to evaluate the performance of changed "hardware" rather than just varying input values. Such testing technique is known in software engineering as fault injection (Hsueh et al. 1997).

Computer models of phenomena that exhibit probabilistic properties can be tested for internal consistency reliability through multiple replications (Sargent 1991). In the construction domain, such phenomena include e.g., activity durations (Lee 2005), cost estimates and bids (Chua et al. 2001), and soil conditions (Karam et al. 2007). A direct relationship exists with the quality of the input data, which are replaced by input distributions (AbouRizk et al. 1994) from which samples are drawn in a quasi-random manner. In each repetition, or run, the specific realization of the probabilistic variable is created via a random number generator. Isidore and Back (2002) analyzed risk from both schedule and cost with "nested" repetitions of "25 times using 100 runs per simulation." A direct extension of the replication approach is the widely used Monte Carlo method, which—more powerful than sensitivity analysis—seeks to quantify the probability parameters of the output in cases when the probabilistic input is too complex for a direct, deterministic analysis (Chantaravarapan et al. 2004).

Opportunities

The construction research community recognizes the importance of validation of research efforts as well as the challenges associated within the construction domain. Indeed, the National Science Foundation (NSF), the National Institute of Standards and Technology (NIST), and the Fully Integrated and Automated Technologies for Construction (FIATECH) industry consortium, are collaborating to foster the dialog about construction *test-beds*. Test-beds include items such as physical laboratories, mathematical models, algorithms, computer simulations, and databases, as well as their integration into sophisticated systems. They allow for rigorous and replicable testing to realistically and comprehensively evaluate the impact of research endeavors.

FIATECH's Intelligent and Automated Construction Job Site (IACJS) test-bed was identified by its *Capital Projects Technology Roadmap* development team as one of the strategies needed to help achieve the vision of the construction job site of the future. The objective is to "focus on the design, development, deployment, and use of an advanced set of test facilities that will provide the first major tests of the IACJS test-bed concepts. This test-bed will provide realistic environments in a laboratory setting from which proposed and developed IAJCS technologies can be tested for potential use on actual construction sites" (FIATECH 2004). FIATECH recognizes that having this capability will accelerate industry adoption of new ideas that could improve safety, cut costs, and reduce construction schedules. The IACJS test-bed

is considered as a critical element in the process of developing and verifying metrics and tools to attain improvements in construction productivity. FIATECH and NIST are currently collaborating to (1) identify a draft set of requirements (capabilities) for an IACJS test-bed; (2) identify an initial set of high-impact target technologies, processes, or methods to evaluate within the IACJS test-bed; and (3) identify industry partners interested in using developed test-bed capabilities.

NSF and NIST are also working together toward the development of NSF-NIST Advanced Building Infrastructure Test-Beds (NABIT). "There is a variety of ongoing research related to intelligent infrastructure, building commissioning, building information integration systems that have special needs related to their verification and validation. These system-level projects employing advanced information and communication technologies (ICT) and methods require sophisticated test-beds to realistically and comprehensively evaluate the impact and process improvements brought about by the products of that research. The use of test-beds as a means of verifying and validating this ICT-related research is critical to the successful evolution of this field of inquiry, both among researchers as well as between researchers and the user community. There are many NSF-funded projects that could benefit greatly from awareness of and access to existing test-beds through which they could verify and validate their research. There are many potential resources at NIST that would make excellent test-beds for such research. What is needed is a vehicle to bring together institutional needs of students and faculty, and potential verification and validation resources (facilities, equipment, and data) developed by NIST staff" (Akin and Garrett 2006).

As a result of the NABIT initiative, a database of test-beds is currently being developed under the umbrella of the Virtual Community of Construction Scholars and Practitioners ⟨ www.cscholars.washington.edu ⟩. It is partially sponsored by the NSF and the objective of its database is to provide an avenue for construction researchers to easily identify potential test-beds that are available to validate their research. Once this database is populated, it will include information such as the type of test-bed, a brief description, its past uses, and a contact person to gather additional information and gain access to the test-bed. Government agencies and members of the construction community will be invited to upload information into the database.

Construction researchers must realize that models and tools created in the performance of their own research may very well serve as test-beds for other research, even when they were not developed with that purpose. For example, Rojas and Mukherjee (2006) developed a situational simulation termed "Virtual Coach" to support the educational process of future construction project managers. This gamelike environment was funded by the U.S. Department of Education, given its potential impact on learning paradigms. However, as a result of this effort, the dynamics of the construction phase of commercial projects have been modeled and the system now also supports significant user interaction as well as the creation of new simulations by nonprogramers. Therefore, once face validity of this model is established, it can serve as a test-bed to evaluate various hypotheses. One could imagine the performance of a study about information timeliness by testing the sensitivity of the system to changes in the delivery time of information to decision-makers. Some test groups may receive real-time data, while others may experience delays ranging from a few days to a couple of weeks. The advantage of such a scheme is that test groups would be exposed to the same simulated construction environment and would experience the same simulation

JOURNAL OF CONSTRUCTION ENGINEERING AND MANAGEMENT © ASCE / JANUARY 2010 / **133**

J. Constr. Eng. Manage., 2010, 136(1): 127-135

events. Indeed, this may create a virtual laboratory where the performance of each group can be evaluated by analyzing their execution in terms of schedule and budget.

Conclusion

Validation of the research methodology and its results is a fundamental element of the process of scholarly endeavor. Approaches used for construction engineering and management research have included experiments, surveys and observational studies, modeling and simulation, theory building, case studies, and various subtypes thereof. Some studies use more than one approach. A particular contribution of this paper has been reviewing different types of validation using examples of studies, analyzing the specific challenges that were found to be significant, and presenting how they were successfully overcome in each case. Another contribution has been describing new opportunities for research validation that are emerging at the horizon as well as ongoing collaborative efforts to enhance the access of construction researchers to validation tools. This paper has increased the awareness of the paramount role that validation techniques play in scholarly work by providing readers with recommendations to properly validate their own research efforts.

Acknowledgments

References

AbouRizk, S. M., Halpin, D. W., and Wilson, J. R. (1994). "Fitting beta distributions based on sample data." *J. Constr. Eng. Manage.*, 120(2), 288–305.

Akin, O., and Garrett, J. H. (2006). "Final report: NSF-NIST workshop on advanced building infrastructure testbed (NABIT)." Division of Civil and Mechanical Systems, National Science Foundation, Arlington, Va.

Anderson, S. D., Patil, S. S., Gibson, G. E., and Sullivan, G. R. (2004). "Owner-contractor work structures: Process approach." *J. Constr. Eng. Manage.*, 130(5), 680–690.

Babbie, E. (1990). *Survey research methods*, Wadsworth, Belmont, Calif., ⟨http://www.cengage.com/wadsworth⟩.

Bartlett, J. E., Kotrlik, J. W., and Higgins, C. C. (2001). "Organizational research: Determining appropriate sample size in survey research." *Business Education Index*, 19(1), 43–50.

Barton, R. R., and Schruben, L. W. (2001). "Resampling methods for input modeling." *Proc., 2001 Winter Simulation Conf.*, B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, eds., IEEE, Piscataway, N.J., 372–378.

Blayse, A. M., and Manley, J. (2004). "Key influences on construction innovation." *Constr. Innovation*, 4(3), 143–154.

Chantaravarapan, S., Gunal, A., and Williams, E. J. (2004). "On using Monte Carlo methods for scheduling." *Proc., 2004 Winter Simulation Conf.*, R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds., IEEE, Piscataway, N.J., 1870–1975.

Chua, D. K. H., Li, D. Z., and Chan, W. T. (2001). "Case-based reasoning approach in bid decision making." *J. Constr. Eng. Manage.*, 127(1), 35–45.

Col Debella, D., and Ries, R. (2006). "Construction delivery systems: A comparative analysis of their performance within school districts." *J. Constr. Eng. Manage.*, 132(11), 1131–1138.

Cresswell, J. W. (1994). *Research design: Qualitative and quantitative approaches*, SAGE, Thousand Oaks, Calif., ⟨http://www.sagepub.com/home.nav⟩.

Cronbach, L. J. (1951). "Coefficient alpha and the internal structure of tests." *Psychometrika*, 16(3), 297–335.

Davis, K. A. (2004). "Information technology change in the architecture, engineering, and construction industry: An investigation of individuals' resistance." Ph.D. dissertation, Virginia Polytechnic Institute and State Univ., Blacksburg, Va.

Davis, K. A., and Songer, A. D. (2003). "Individuals' resistance to technological change in the AEC industry." *Proc., 2003 Construction Research Congress*, K. R. Molenaar and P. S. Chinowsky, eds., ASCE, Reston, Va., 8.

del Cano, A., and de la Cruz, M. P. (2002). "Integrated methodology for project risk management." *J. Constr. Eng. Manage.*, 128(6), 473–485.

Elazouni, A. M., and Metwally, F. G. (2007). "Expanding finance-based scheduling to devise overall-optimized project schedules." *J. Constr. Eng. Manage.*, 133(1), 86–90.

El-Diraby, T. E., and O'Connor, J. T. (2004). "Lessons learned in designing research methodology in field-based construction research." *J. Profl. Issues Eng. Educ. Pract.*, 130(2), 109–114.

Elmaghraby, S. E. (2000). "On criticality and sensitivity in activity networks." *Eur. J. Oper. Res.*, 127, 220–238.

FIATECH. (2004). *Capital Projects Technology Roadmapping Initiative: Element 4 Tactical Plan, Intelligent and Automated Construction Job Site (IACJS)*, Fully Integrated and Automated Technologies for Construction, Austin, Tex.

Gay, L. R., Mills, G. E., and Ariasian, P. (2006). *Educational research: Competencies for analysis and applications*, 8th Ed., Prentice-Hall, Upper Saddle River, N.J.

Gibson, G. E., Kaczmarowski, J. H., and Lore, H. E. (1995). "Preproject-planning process for capital facilities." *J. Constr. Eng. Manage.*, 121(3), 312–318.

González-Quevedo, A., AbouRizk, S. M., Iseley, D. T., and Halpin, D. W. (1993). "Comparison of two simulation methodologies in construction." *J. Constr. Eng. Manage.*, 119(3), 573–589.

González-Quevedo, A. A. (1995). "Sensitivity analysis of construction simulation." *Proc., 2nd Congress on Computing in Civil Engineering*, J. P. Mohsen, ed., ASCE, Reston, Va., 1427–1434.

Gujarati, D. N. (1995). *Basic econometrics*, 3rd Ed., McGraw-Hill, New York.

Hsueh, M.-C., Tsai, T. K., and Iyer, R. K. (1997). "Fault injection techniques and tools." *Computer*, 30(4), 75–82.

Ioannou, P. G., and Martínez, J. C. (1996). "Animation of complex construction simulation models." *Proc., 3rd Congress on Computing in Civil Engineering*, J. A. Vanegas and P. S. Chinowsky, eds., ASCE, Reston, Va., 620–626.

Isidore, L. J., and Back, W. E. (2002). "Multiple simulation analysis for probabilistic cost and schedule integration." *J. Constr. Eng. Manage.*, 128(3), 211–219.

Jaafari, A., Manivong, K. K., and Chaaya, M. (2001). "VIRCON: Interactive system for teaching construction management." *J. Constr. Eng. Manage.*, 127(1), 66–75.

Kallantzis, A., Soldatos, J., and Lambropoulos, S. (2007). "Linear versus network scheduling: A critical path comparison." *J. Constr. Eng. Manage.*, 133(7), 483–491.

Kamat, V. R., and Martínez, J. C. (2003). "Interactive discrete-event simulation of construction processes in dynamic immersive 3D virtual worlds." *Proc., 2003 Conf. on Construction Applications of Virtual Reality*, W. Y. Thabet and D. A. Bowman, eds., Virginia Polytechnic Institute and State University, Blacksburg, Va., 197–201.

Karam, K. S., Karam, J. S., and Einstein, H. H. (2007). "Decision analysis applied to tunnel exploration planning. I: Principles and case study." *J. Constr. Eng. Manage.*, 133(5), 344–353.

Kelton, W. D., and Barton, R. R. (2003). "Experimental design for simulation." *Proc., 2003 Winter Simulation Conf.*, S. Chick, P. J. Sánchez,

D. Ferrin, and D. J. Morrice, eds., IEEE, Piscataway, N.J., 59–65.

Kolisch, R., and Sprecher, A. (1996). "PSPLIB—A project scheduling problem library." *Eur. J. Oper. Res.*, 96(1), 205–216.

Krejcie, R. V., and Morgan, D. W. (1970). "Determining sample size for research activities." *Educ. Psychol. Meas.*, 30, 607–610.

Law, A. M. (2007). *Simulation modeling and analysis*, 4th Ed., McGraw-Hill, New York.

Lee, D.-E. (2005). "Probability of project completion using stochastic project scheduling simulation." *J. Constr. Eng. Manage.*, 131(3), 310–318.

Leedy, P. D., and Ormrod, J. E. (2001). *Practical research planning and design*, 7th Ed., Prentice-Hall, Upper Saddle River, N.J.

Ling, Y. Y. (2002). "Model for predicting performance of architects and engineers." *J. Constr. Eng. Manage.*, 128(5), 446–455.

Lucko, G. (2007). "Flexible modeling of linear schedules for integrated mathematical analysis." *Proc., 2007 Winter Simulation Conf.*, S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds., IEEE, Piscataway, N.J.

Lucko, G., Anderson-Cook, C. M., and Vorster, M. C. (2006). "Statistical considerations for predicting residual value of heavy equipment." *J. Constr. Eng. Manage.*, 132(7), 723–732.

Menches, C. L., and Hanna, A. S. (2006). "Quantitative measurement of successful performance from the project manager's perspective." *J. Constr. Eng. Manage.*, 132(12), 1284–1293.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2001). *Introduction to linear regression analysis*, 3rd Ed., Wiley, New York.

Navon, R., and Berkovich, O. (2005). "Development and on-site evaluation of an automated materials management and control model." *J. Constr. Eng. Manage.*, 131(12), 1328–1336.

Neuman, W. L. (2000). *Social research methods: Qualitative and quantitative approaches*, 4th Ed., Allyn and Bacon, Needham Heights, Mass.

Ozorovskaja, R., Voordijk, J. T., and Wilderom, C. P. M. (2007). "Leadership and cultures of Lithuanian and Dutch construction firms." *J. Constr. Eng. Manage.*, 133(11), 900–911.

Price, A. D. F., and Chahal, K. (2006). "A strategic framework for change management." *Constr. Manage. Econom.*, 24(3), 237–251.

Rekapalli, P. V., and Martínez, J. C. (2007). "A message-based architecture to enable runtime user interaction on concurrent simulation-animations of construction operations." *Proc., 2007 Winter Simulation Conf.*, S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew,

and R. R. Barton, eds., IEEE, Piscataway, N.J., 2028–2031.

Rojas, E. M., and Aramvareekul, P. (2003). "Is construction labor productivity really declining?" *J. Constr. Eng. Manage.*, 129(1), 41–46.

Rojas, E. M., and Dossick, C. S. (2008). "Developing a state-of-the-art facility to support construction research and education: A case study." *J. Profl. Issues Eng. Educ. Pract.*, 134(1), 67–74.

Rojas, E. M., and Kell, I. (2008). "Comparative analysis of project delivery systems cost performance in Pacific Northwest public schools." *J. Constr. Eng. Manage.*, 134(6), 387–397.

Rojas, E. M., and Mukherjee, A. (2006). "A multi-agent framework for general-purpose situational simulations in the construction management domain." *J. Comput. Civ. Eng.*, 20(3), 165–176.

Roschke, P. N. (1994). "Validation of knowledge-based system with multiple bridge rail experts." *J. Transp. Eng.*, 120(5), 787–806.

Sargent, R. G. (1991). "Simulation model verification and validation." *Proc., 1991 Winter Simulation Conf.*, W. D. Kelton, G. M. Clark, and B. L. Nelson, eds., IEEE, Piscataway, N.J., 37–47.

Snee, R. D. (1977). "Validation of regression models: Methods and examples." *Technometrics*, 19(4), 415–428.

Telem, D., Shapira, A., Goren, Y. D., and Schexnayder, C. J. (2006). "Moving a reinforced-concrete building: Case study." *J. Constr. Eng. Manage.*, 132(2), 115–124.

Thomas, H. R., Riley, D. R., and Sanvido, V. E. (1999). "Loss of labor productivity due to delivery methods and weather." *J. Constr. Eng. Manage.*, 125(1), 39–46.

Vanhoucke, M. (2006). "Work continuity constraints in project scheduling." *J. Constr. Eng. Manage.*, 132(1), 14–25.

Wu, J. Y. (2005). "Soil-based flowable fill for pipeline construction." *Proc., 2005 Pipeline Division Specialty Conf.*, C. Vipulanandan and R. Ortega, eds., ASCE, Reston, Va., 925–938.

Yu, A. T. W., Shen, Q., Kelly, J., and Hunter, K. (2006). "Investigation of critical success factors in construction project briefing by way of content analysis." *J. Constr. Eng. Manage.*, 132(11), 1178–1186.

Zayed, T. M., and Halpin, D. W. (2005a). "Pile construction productivity assessment." *J. Constr. Eng. Manage.*, 131(6), 705–714.

Zayed, T. M., and Halpin, D. W. (2005b). "Productivity and cost regression models for pile construction." *J. Constr. Eng. Manage.*, 131(7), 779–789.

Zhang, H., Shi, J. J., and Tam, C. M. (2002). "Iconic animation for activity-based construction simulation." *J. Comput. Civ. Eng.*, 16(3), 157–164.

JOURNAL OF CONSTRUCTION ENGINEERING AND MANAGEMENT © ASCE / JANUARY 2010 / **135**

J. Constr. Eng. Manage., 2010, 136(1): 127-135