An Empirical Approach to Design Metrics and Judgments

Dewayne E. Perry
Center for Advanced Research In Software Engineering (UT ARISE)
The University of Texas at Austin

Abstract

Scientific evaluations and comparisons of designs require a mature set of design constructs and observable measures. Currently we do not have that maturity. The requirements are defined for an experimental basis for developing design constructs and observable measures that can then be used to support scientific judgments about designs and design methods and techniques.

1. Introduction

Design is at the core of any engineering discipline and software engineering is no exception. How good the designs are critically determine the effectiveness and viability of the software systems we build, market and evolve. Moreover, the techniques and methods for architectural detailed design are critical in creating architectures and designs that meet the constraints imposed by user requirements, marketing demands, company goals, and project constraints.

If we are to advance significantly in the areas of creating architectural and detailed designs that meet their intended constraints using appropriate methods and techniques, we must be able

- to determine which methods and techniques work best relative to the desired design domain and architectural and design constraints, and
- to measure and evaluate architectural and design alternatives both quantitatively and qualitatively in order to make sound judgments about these alternatives.

Currently, we have neither the knowledge nor mechanisms to determine the most appropriate methods and techniques, nor do we have the measurements needed to evaluate and compare the possible alternatives with any confidence. We are still at the stage of maturity that is nearer to that of a craft than an engineering discipline. While we can often tell from experience that design X is better than design Y, we do these judgments from the standpoint of an art critic, not from the standpoint of an engineer appealing to well understand measures derived from a well established theoretical basis.

In the following I will first discuss the problem of measurement and evaluation. Once the basis for developing meaningful measures for evaluating designs and design methods and techniques has been resolved then we can discuss the problems of evaluating design methods and techniques and their appropriateness for various domains and constraints.

2. Design Metrics

In the design of any experiment, one of the first critical issues that arises is that of *construct validity*.[1] In formulating our research hypotheses we may express them either in terms of

---

1. *Discussions of construct validity can be found in any number of texts of behavioral science research. The ones I have used in preparing this white paper are 1) Julian Meltzoff, Critical Thinking About Research: Psychology and Related Fields, American Psychological Association, Washington DC, 1997; and 2) Robert Rosenthal and Ralph Rosnow, Essentials of Behavioral Research: Methods and Data Analysis, Second Edition, McGraw-Hill Series in Psychology, McGraw-Hill, 1991.*

abstractions (referred to as *constructs*) or in terms of *observable measures*. If we formulate them in terms of abstractions or constructs, we must eventually represent these abstractions by observable measures that can be used in the design of the experiment. If we formulate the hypotheses in terms of observable measures, we still must relate them back to the appropriate abstractions or constructs in order to frame the research questions in the context of some theoretical structure.

So, the critical part at this stage of experimental work is to determine the required constructs and to make sure that the observable measures represent these constructs properly — ie, satisfy the demands of *construct validity*. To do that we must satisfy two different problems:

- *representation* (or translation) validity, and
- *observation* (or criterion) validity.

2.1 Representation Validity

Representation validity is concerned about how well the constructs or abstractions translate into observable measures. There are two primary questions to be answered.

- Do the subconstructs properly define the construct (if you break up the main abstractions into smaller abstractions or definitions)?
- Do the observations properly interpret, measure, or test the constructs?

One way to argue positively, albeit a very weak argument, is to claim *face validity* for the construct/observable relationship. Basically this is making the following claim: on the face of it, it seems like a good translation. The weakness of this argument can be strengthened by a consensus of experts.

Another way to argue positively is to claim *content validity* for the construct/observable relationship. To do this one must check the operationalization against the relevant content domain for the construct: to extent to which the tests (ie, the observable measures) measure the content of the subject being tested — ie, that all the important content areas are covered adequately.

2.2 Observation Validity

Having decided on the appropriate design constructs and and the fact that their observational representatives are adequate translations of those constructs, we now must focus our attention on the quality of the observable measures themselves.

There are four basic requirements on observable measures:

- *predictive validity,*
- *concurrent validity,*
- *convergent validity*, and
- *discriminant validity*.

Predictive validity means that the observed measure predicts what it should predict and nothing else. For example, tests of college aptitude are assessed according to how well they predict success (grades, graduation, etc) in college.

Concurrent validity means that the observed measure correlates highly with an established set of measures. For example, shorter forms of tests are evaluated for their concurrent validity using the longer form of tests.

Convergent validity means that the observed measure correlates highly with other observable measures for the same construct. However, the utility of such a measure is not that it duplicates some other measure but that it is a new way of distinguishing a particular trait while correlating well with other similar measures.

Divergent validity means that the observable measure distinguishes between two groups that differ on the trait in question. The utility of such a measure is that it is able to differentiate among similar groups along the lines that are critical for that measure.

Having satisfied these requirements, we then have an observable measure that is predictive, intentional, well-behaved and discriminating. One additional useful characteristic is that it should be invulnerable to observer biases — that is, the use of the measure is both *reliable* and *stable* across a wide ranges of uses and users.

2.3 What Do We Do Now?

Now comes the hard work! The design metrics that have been proposed have neither arguments about their representation validity, nor a consensus among the community. Indeed, some design constructs such as cohesion seem to defy translation into any form of observable measure. A very useful metaphor no doubt, but very difficult to translate successfully into either a set of subconstructs or into a set of observable measures.

Nor have the various observable measures proposed routinely and frequently been subjected to the scrutiny that is required of measures in the behavioral sciences. We have done very little to establish the predictive, concurrent, convergent, and discriminant validity of these measures. Furthermore, we have no data on their reliability or stability, whereas the behavioral scientists have not only established these issues of validity they have established the reliability and stability for them as well.[2]

Thus we have a very large amount of work to do:

- First, serious work has to be done determining useful design constructs and their decomposition into subconstructs. We need to establish a consensus on what abstractions are critical to our being able to evaluate and compare designs.
- Second, serious work has then to be done to find appropriate observable measures and establish their representational validity.
- Thirdly, having established their representational validity, we must the establish there observational or criterion validity and their reliability and stability as observable measures.

---

2. *There is a 6 volume set by Goldman et al, Directory of Unpublished Experimental Measures, published by the American Psychological Association, 1995-96, and a series of yearbooks edited by Buros, Mental Measurements Yearbook, that enable behavioral scientists to evaluate the reliability and stability of useful observable measures.*

All in all, a very large amount of work for a very large set of experimental researchers. And it must be done if we are ever to leave the age of the art critic and enter the age of scientific evaluation.

3. Design Judgments

Once we have a set of design constructs and metrics we can then begin to take design evaluation beyond the current state.

What we now can do at best is to justify the design after the system has been completely built by appealing to how well it performs its various function, how usable it is, how easy it is to evolve, how successful it is in the market place etc. We take on faith that what we have done is better than what we have not done. Or we appeal to our experience and what has worked and what hasn't.

With a proper set of design constructs and measures we can objectively evaluate and determine who well a particular design meets particular constraints or whether one design is better than another with respect to specific design criteria.

Until then, we remain art critics, judging on the basis of internalized standards and personal opinions.

4. Postscript: Domain Specific Knowledge

There is considerable support empirically for domain specific approaches to building software systems. Curtis et al[3] noticed the ''thin spread of application knowledge'' in studying the design process. Perry and Steig[4] found that a substantial proportion of the faults found had as their underlying cause the lack of domain and system knowledge. This observation was further strengthened in Leszak et al[5] root cause analysis study of a particular release of a network product. Again, the dominant underlying root cause of the software faults found was the lack of domain and system knowledge.

Given this consistent story of domain and system specific knowledge as a major cause of software system faults, design methods, techniques and technologies that focus on domain specific solutions has the potential for reducing or eliminating a significant set of software faults.

---

3. *Curtis, et al. "A Field Study of the Software Design Process for Large Systems", CACM 31:11 (November 1988), 1268-1287.*
4. *Dewayne E. Perry and Carol S.Steig, "Software Faults in Evolving a Large, Real-Time System: a Case Study", 4th European Software Engineering Conference -- ESEC93, Garmisch, Germany, September 1993.*
5. *Marek Leszak, Dewayne E Perry and Dieter Stoll. "A Case in Root Cause Defect Analysis", International Conference on Software Engineering 2000, Limerick Ireland, June 2000. An expanded version will appear in the Journal of Systems and Software.*