

The Roots of CWS

David J. Weiss

History is bunk.

Henry Ford

The genesis of CWS can be traced back to 1966, when James Shanteau (JCS) and David Weiss (DJW) were in their first year of graduate school at UCSD in La Jolla, CA. They shared an office and a mentor, Norman H. Anderson, who instilled in them a love of analysis of variance (ANOVA).

Fast-forward now to 1979. DJW was a faculty member at California State University, Los Angeles. In those glorious times, the University was able to indulge in some wondrous courses. One of DJW's colleagues, David Fitzpatrick, a researcher in perception and self-professed wine expert, planned to offer an extension course entitled "Sensory Appreciation of Wine." As one might expect, this instructional adventure called for some mocking.

Eventually, a serious evaluation was proposed. DJW, whose early research work was in the area of psychophysical judgment, proposed to test Fitz's students objectively. Superficially, this was a straightforward exercise. The class members would be asked to blindly rate the likability of nine wine samples, taken from bottles that varied in price, at the beginning and end of the course. The problem, though, was that it was not clear how to evaluate the tasting responses, since there were no right answers. Undaunted, and unconstrained by any knowledge of the literature on expertise (or even the recognition that this was an issue of expertise), DJW constructed two ways to resolve this deficit. One solution was to regard the instructor's ratings as correct (we would now call this a

“gold standard”). An “accuracy” measure was computed for each student’s initial and final ratings that consisted of the root mean square of the deviations from the instructor’s ratings (re-inventing the Brier score). The correlation between each student’s ratings and the instructor’s ratings was also examined. These two measures are associated but not identical, and both yielded the result that the students did not improve over the course.

More important to the development of CWS was the other solution DJW constructed to deal with the absence of correct answers. Drawing upon his fondness for analysis of variance, he proposed to assess “discriminating power”, the ability of the judge to assign reliably different scores to different wines. Reliability, which we now call consistency, examined the extent of the differences between a given person’s ratings of the same stimulus when it was presented more than once; this construct was captured by the variance within cells. Discrimination meant that different wines should inspire different ratings, and was captured by the variance between cells. Using a single-S design with two scores per stimulus per subject, an F-ratio could be constructed from each student’s initial ratings. The validity of this measure was demonstrated by the instructor’s F-ratio, 10.01, being higher than the average student’s F-ratio, 5.53. Foolishly, DJW omitted to collect F-ratios for the final ratings, because his rather limited goal had been to see whether students with superior discriminating power were better able to learn to emulate the instructor (they weren’t).

The wine-tasting results were presented at a Mathematical Psychology meeting (Weiss, 1980) to the usual small, yawning audience (JCS was not in the “crowd”), and they lay fallow for several years. During the early 1980’s, DJW and JCS collaborated on a statistical computer program (Weiss & Shanteau, 1982). The pairing was a one-shot

arrangement; both viewed the algorithm as secondary to their substantive research interests. They did not know that they had been working on a common problem. JCS had begun to swing his research in the direction of expertise, but DJW did not realize that his little study of “training” was worth mentioning.

Having fallen in love with the BASIC computer language during the time of that collaborative effort, DJW came upon an obscure statistical procedure advanced by Schumann and Bradley (1959), in which the authors proposed to determine whether one F-ratio was significantly greater than another. DJW eventually published a program that implements the procedure (Weiss, 1985). The empirical illustration Schumann and Bradley used was an experiment by William Cochran (1943), a prominent statistician, who compared the efficacy of response instruments using an F-ratio criterion. This criterion was identical to the “discriminating power” measure. Cochran’s criterion was applied to instruments rather than to people, but his analysis used an F-ratio in the same way DJW had assessed the wine tasters. DJW consoled himself with the thought that, in those days before computerized literature searches, relevant references were often missed. The Schumann-Bradley procedure eventually became a standard CWS tool for comparing sets of judgments.

Fast-forward again to the mid-1990s. JCS had become an expert on experts, and was disenchanted with the subjectivity in the field. The only serious effort to assess experts objectively was that of Einhorn (1972, 1974), whose proposal didn’t really catch on. But JCS admired Einhorn’s attempt, and constructed a table showing observed reliability (within-individual correlation) and observed consensus (between-individual correlation) for experts from several domains. This table appeared, in several

incarnations, at a host of conference presentations and was eventually published (Shanteau, 2001). JCS and members of his various audiences attempted to find structure in the patterns of correlations, but this exercise rivals interpreting the Kabbala. We now believe that the table piqued interest but was ultimately frustrating because it was half-right. Reliability is a crucial component of expertise, but consensus is not (Weiss & Shanteau, 2004). As an audience member, DJW played the structure-seeking game too, but with little success. He recalled to JCS his work on what he now knew was the topic of expertise, and showed him the notes from his 1980 presentation. JCS realized there was promise in the approach, and the two began to develop the index and sought applications.

DJW had seen his measure of discriminating power as an analogy, arguing that an expert judge should be like a measuring instrument. The psychophysics background directed his thinking. Instruments assign different numbers of different objects, and they do so reliably. It wasn't until JCS began to reanalyze previously collected data on various grades of experts that the pair began to realize that the ratio captured a property that characterized people. The experts had produced higher F-ratios than the novices in several studies of judgmental skill, although the original data had not been analyzed in this way.

JCS realized that an F-ratio was only one example of the general index that could be constructed. Any measure of discrimination could be compared to any measure of consistency. And he saw that the F-table was irrelevant, in that the observed F-ratio was tied to the stimuli used and could be interpreted only on a relative basis. Sensing the importance of a new name for the index, JCS came up with "CWS", to honor Cochran

(C) for his historical precedence and to acknowledge the creators, Weiss (W) and Shanteau (S).

JCS also established an important context for application of the new index. Working with Julia Pounds, one of his former Ph.D. students who became a researcher with the Federal Aviation Administration, he demonstrated the value of CWS in examining the expertise displayed by air traffic controllers. This application provided a funding basis that supported the development of CWS. The funding allowed JCS to provide financial support for some wonderful Kansas State graduate students, among them Brian Friel, John Raacke, and especially Rick Thomas, who have contributed significantly to the empirical work that validates the approach. Shawn Farris revamped DJW's primitive website into the professional vehicle that now exhibits our collected works.

The empirical work began to show that CWS captured expertise in performance as well as judgmental tasks. This was remarkable, because the index was purported to work at the level of judgment. One might anticipate that there would be quite a bit of psychological processing between judgment and execution of a skilled behavior. Apparently, for many tasks, the overlain behaviors do not introduce nonlinearities that would mask the judgmental component. When the results that Thomas brought in showed that CWS scores predicted such correlates of good performance as task complexity and amount of practice, the CWS creators became more than proponents, they became true believers in the ability of the ratio to capture expertise.

The aspect of expertise CWS that captures is what a professional needs in everyday practice. The approach does not address the issue of originality. The people

whom we as a culture value most, the Nobel laureates and creative artists, exhibit expertise of a kind that we do not consider at all. CWS does not purport to be the whole story. (But we can help identify those who do the best job of judging to whom the Prize should be awarded...)

An important moment came in 1999, when Ward Edwards heard DJW and JCS present CWS ideas. Edwards, who can perhaps be regarded as the founding father of psychologically based decision-making research, was effusive in his praise. At the same time, he asked penetrating questions. This encouragement provided a good deal of momentum to the CWS team.

Team Chemistry

Prior to the CWS project, DJW and JCS have had long careers as professors and researchers. Like most of their counterparts, they had usually conducted research as lone wolves, working either alone or with student collaborators. Student collaborators inevitably have a subordinate status, even when they are no longer students. The power differential resulting from a disparity of experience translates into deference, which deprives the group of pointed contributions from the junior members.

Collaboration with a peer is different. DJW and JCS do not mind criticizing one another's ideas, always in a congenial way, often in a humorous vein, but never holding back because of perceived fear of retribution. This makes for lively and productive interchange, with the project receiving the full benefit of each member's insights. They make a point of modeling this behavior in front of the student members of the team. It is not easy for a student to suggest that a professor is wrong, even though there may have been lots of evidence that such things do happen. Students have a tough time grasping

that we as professors feel pride when caught by our carefully molded students. We try to state that explicitly.

Two issues illustrate the kind of challenging debates we had in those heady days when we were developing the index. Rick Thomas proposed that we incorporate time into the CWS index for cases when time pressure is imposed on the expert. The notion appealed; in contexts such as surgery or financial trading, fast thinking and fast execution are valued. DJW, who teaches a class in sexuality, offered a counterexample; but there was general agreement that faster is usually better. We built an CWS-T(ime) formula that incorporated a time component, but the problem was that that time is measured in different units from the other elements. How much time is equivalent to a given increment in discrimination or in consistency? If one candidate discriminates more but another is faster, who gets the higher CWS-T score? We could not see how to resolve that dilemma without incorporating expert opinion to determine the tradeoff, and that violates the spirit of our enterprise. We have not abandoned thoughts of time pressure; but we now view it as a task variable, akin to stimulus complexity, rather than as a core element in our index.

Another debate concerned the form of the index. The CWS ratio is unbounded, and we found empirical cases in which the obtained values for some candidates were hundreds of times larger than those of their colleagues. This wide range bothered some of our audience members. CWS may appear somewhat easier to interpret if we standardize the ratio into a proportion-of-variance-like quantity akin to η^2 . A proportionate representation (CWS_P), although monotonically related to the ratio as originally defined, has the advantage that values range from zero to one.

$$CWS_P = \frac{\text{Discrimination}}{\text{Discrimination} + \text{Inconsistency}}$$

A value of .5 means that inconsistency is equal to discrimination; no expertise has been demonstrated. Therefore, values below .5 are sampling errors. We have been reluctant to recast the index in this proportionate form, because the standardization makes it seductively easy to compare values that ought not be compared. Comparisons are meaningful only when stimuli are identical and the same response measures have been used.

The wide range in observed CWS values generally arises because of disparities in consistency. Consistency can result when judges remember their responses to repeated stimuli that they recognize. Although we might urge candidates to judge each stimulus independently, we cannot easily control whether they do so. Stimuli that are identifiable, perhaps because they include names, are especially susceptible to mnemonic strategies. Self-presentation concerns may drive people to appear consistent, since it is well-known that experts ought to be consistent (and it will be even better known as CWS is popularized!) Is the wide range a problem for the index? As Edwards observed, a component of expertise is the ability to connect previous stimuli and responses to current ones. Accordingly, peculiarities in the distribution of observed CWS scores are not scale anomalies, but rather have psychological meaning. We should not try to transform them away, but instead work toward trying to understand them.

Email has made collaboration across campuses easy. However, we still find much value in face-to-face meetings, and are always pleased by the new ideas that flow from such gatherings.

The Didactic Power of Good Examples

Presenting a novel approach to an old problem has its challenges. As we simultaneously developed the theory and applications of CWS, DJW and JCS spoke at conferences and conducted workshops describing their new toy. We anticipated that people would be dazzled by its brilliance. Unfortunately, we encountered a fair number of people who didn't get it. Naturally, we ascribed their difficulties to cognitive limitations, biases, etc. It took us a while to see that we had obscured the practical value of CWS with poorly chosen examples.

Our favorite illustration uses medical judgments collected by Skånér, Strender, and Bring (1998). JCS drew a graph showing four physicians' judgments of the probability of heart failure for five patients. From visual inspection alone, it is obvious that one of the doctors is both discriminating and consistent, while the judgments of the other three are deficient in one or both of those characteristics that we hold dear. Most audience members nod when we present the graph. Occasionally, some thoughtful members of the audience realize that strategic elements as well as pure diagnostic skill might enter into the judgments. This insight leads to lively discussion, and those involved are hooked.

Because we fell in love with that illustration, we got carried away and used a related case in our introduction. We asked the audience members to imagine themselves in a hypothetical situation in which they're in a city far from home and begin to experience chest pains. They want to consult a physician, but are faced with the problem of identifying a good one. How to solve that problem, we asked. We provided the answer to that rhetorical question – use CWS to assess the doctors.

Of course, that is a ridiculous solution. What any sensible person would do in that scenario is to try to find someone with local knowledge and ask for a recommendation. One might look for a nearby medical school. Our example confused people so badly that some never got untracked. We needed to provide different examples at the introductory point.

We also needed to distinguish carefully between experts and expertise. Experts are people, and expertise describes ability to do a specific task. An expert is someone who exhibits expertise on a set of relevant tasks. Separating these constructs made our presentation much more comprehensible. Now when we present CWS to a new audience, we often encounter the reaction that it's all obvious. This is the most flattering evaluation possible.

Overcoming Reluctance to be Assessed

Another of our early disappointments was that real experts didn't seem to be eager to have their expertise assessed with this new objective technique. Oh, we could get experts to participate if we were planning to contrast their performance with that of novices, but a pure study of practicing experts proved difficult to carry out.

It was only when we thought about assessing our expertise as professors that we realized the problem. We view the primary data used to evaluate instructional performance, student ratings, as essentially meaningless. Why does the academic profession maintain this farce? Obviously, because we don't really want to be assessed. So long as the assessment tools are privately acknowledged as invalid, there is no threat. Indeed, why should someone whose expertise has already been established cooperate in the exploration of a new tool that has the potential to question the existing hierarchy?

We realized that experts would be eager to participate in our empirical studies only if there was something in it for them. The paltry financial incentives we offer undergraduates would not impress professionals, but the possibility of improving the field is a goal worthy of some effort. Accordingly, we now advocate studies that explore the effects of training or equipment, of domain complexity or workload when real experts are to be recruited. The promise of objective analysis of such critical variables is a sufficient lure to attract participants whose time is usually more valuable than ours.

References

- Cochran, W. G. (1943). The comparison of different scales of measurement for experimental results. *Annals of Mathematical Statistics, 14*, 205-216.
- Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance, 7*, 86-106.
- Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology, 59*, 562-571.
- Schumann, D. E. W., & Bradley, R. A. (1959). The comparison of the sensitivities of similar experiments: Model II of the analysis of variance. *Biometrics, 15*, 405-416.
- Shanteau, J. (2001). What does it mean when experts disagree? In E. Salas and G. Klein (Eds.), *Linking expertise and naturalistic decision making*. (pp. 229-244). Mahwah, NJ: Erlbaum.
- Skånér, Y., Strender, L., & Bring, J. (1998). How do GPs use clinical information in the judgements of heart failure? *Scandinavian Journal of Primary Health Care, 16*, 95-100.

- Weiss, D. J. (1980, August). *Training the expert judge*. Paper presented at the Mathematical Psychology Meeting, Madison, WI.
- Weiss, D. J. (1985). SCHUBRAD: The comparison of the sensitivities of similar experiments. *Behavior Research Methods, Instrumentation, and Computers*, *17*, 572.
- Weiss, D. J. (1999, February). *The tower of Pisa, the Mona Lisa*. Paper presented at the Bayesian Research Conference, Los Angeles.
- Weiss, D. J., & Shanteau, J. C. (1982). Group-Individual POLYLIN. *Behavior Research Methods and Instrumentation*, *14*, 430.
- Weiss, D. J., & Shanteau, J. (2004). The vice of consensus and the virtue of consistency. In C. Smith, J. Shanteau, & P. Johnson (Eds.), *Psychological investigations of competent decision making*. (pp. 226-240). Cambridge, UK: Cambridge University Press.