

# CWS: A User's Guide

David J. Weiss

James Shanteau

## **What is the CWS approach?**

CWS is the Cochran-Weiss-Shanteau approach to assessing expertise purely from data. The approach is based on the idea that expert judgment involves discrimination – seeing fine gradations among the stimuli – and consistency – evaluating similar stimuli similarly. The approach was inspired by an idea for comparing response instruments suggested by the late statistician William Cochran (1943), and adapted to the domain of expertise by David J. Weiss and James Shanteau.

## **What do you mean by “an expert”?**

Applying the term “expert” to a person is a shorthand description of a set of results rather than a characterization of the person. Talent and training may combine to yield a person we label as expert, but it must be kept in mind that the label is a generalization. It is the behavior that is or is not expert. With CWS, we measure expertise in a specific setting, with specific stimuli and a specific task. Someone who excels in one context may not excel in others that seem similar.

## **What is the CWS index?**

The CWS index is a numerical value that captures the degree of expertise demonstrated in a set of responses. It consists of the ratio of discrimination to inconsistency. Discrimination refers to the candidate's differential evaluation of the various stimuli within a set. Consistency refers to the candidate's evaluation of the same stimuli similarly over time; inconsistency is its complement. The ratio will be large when a candidate discriminates effectively, and will be reduced if the candidate is inconsistent.

$$CWS = \frac{\text{Discrimination}}{\text{Inconsistency}}$$

The rationale for incorporating discrimination and consistency into an index of expertise is that a good measuring instrument, such as a ruler or a thermometer, has these properties. Discrimination and consistency are the building blocks of measurement. Similarly, expertise at its core requires the ability to evaluate the stimuli in one's domain. Note that accuracy is not involved in CWS, as we do not assume any knowledge of correct responses.

## **How are Discrimination and Inconsistency measured?**

For data obtained using interval or ordinal scales, we usually measure both quantities as variances. Discrimination is the variance among averaged responses to different stimuli, while inconsistency is the variance among responses to the same stimulus, averaged across stimuli. The variance, because it entails squaring deviations, has the property that large differences are accentuated. The ratio of the variances follows the F-distribution, and is computed as if it were an F-ratio, but its interpretation is somewhat different. (It is also plausible to measure discrimination and inconsistency with other dispersion measures, such as standard deviation or mean absolute deviation.) Because different stimuli are generally expected to elicit different responses, the measures of discrimination and inconsistency are strictly tied to the set of stimuli actually presented. It is not meaningful to compare measures obtained using different stimulus sets. In this sense, CWS ratios are unlike F-ratios, in that they have no meaning in isolation. The CWS index is scale independent; a linear transformation of the response scale does not alter the ratio.

For data obtained using nominal (categorical) scales, in which responses either agree or disagree with no hint of the extent of disagreement, we have developed measures based on the possible and obtained numbers of (non-)matches across stimuli and replications. These measures are illustrated below in the following example:

Illustration of CWS Index for Nominal Data (four response alternatives)

	Stimulus 1	Stimulus 2	Stimulus 3	Stimulus 4	Stimulus 5
Replicate 1	A	D	B	C	C
Replicate 2	A	B	B	B	B
Replicate 3	A	B	A	B	A
Matches	3	1	1	1	0

For both numerator and denominator, we utilize the proportion of obtained pairwise non-matches to possible matches. In measuring discrimination, a match is evidence of failure to discriminate, so the greater the proportion of observed non-matches, the greater the discrimination. In measuring inconsistency, a match means the response was consistent, so the greater the proportion of observed non-matches, the greater the inconsistency. Expert performance is marked by few matches across columns (stimuli), and many matches within columns (replications). If there are no matches within columns – no consistency at all - the CWS ratio is undefined, but that outcome unambiguously connotes a lack of expertise.

$$\text{CWS Numerator (Discrimination)} = \sum_{\text{Matrix}} \frac{\text{Non – matches across columns}}{\text{Possible matches across columns}}$$

The number of possible matches across columns is most easily calculated by subtracting the number of possible within-column matches from the total number of possible matches. Each response may be matched to any other, so the total number of possible matches is  ${}_{15}C_2$  (= 105). There are  ${}_3C_2$  (= 3) possible matches within each

column, so the number of possible within-column matches is  $5 \times {}_3C_2 (= 15)$ . Therefore, there are 90 possible matches across columns, and 15 possible matches within columns.

In the example above, there were 7 pairs of “A” responses in different columns. “B” was matched 18 times, “C” once, and “D” was not matched at all.

$$\text{Numerator} = \frac{90 - 26}{90} = .711$$

$$\text{CWS Denominator (Inconsistency)} = \sum_{\text{Columns}} \frac{\text{Non - matches within columns}}{\text{Possible matches within columns}}$$

$$\text{Denominator} = \frac{15 - 6}{15} = .60$$

$$\text{CWS Index} = \frac{\text{CWS Numerator}}{\text{CWS Denominator}} = \frac{.711}{.60} = 1.185$$

### **Why is the CWS index set up as a ratio?**

The ratio formulation reflects the idea that a candidate can trade off the two quantities as dictated by the situation. By widening the range of responses used, one can increase discrimination, but only at the cost of decreasing consistency.

Everyone strikes a balance between discrimination and consistency. Performing well in one respect or the other is relatively easy. Someone who can do both at once is behaving expertly.

### **What is the unit of analysis – who is “the candidate”?**

In general, an individual person, a candidate expert, generates a single CWS score for a particular experimental condition. That CWS score may be compared to the score produced by other candidates under identical conditions, or to the CWS score produced by the same candidate under a different experimental condition.

It is also possible for a team to produce a single CWS score, when it is the team's responses to the various stimuli that constitute a data set. In this usage, components of the response from individual team members would not be analyzed separately. The CWS score from one team may be compared to that from another team, or to the CWS score produced by an individual operating alone with the same stimuli.

Designs in which an individual's data constitute the unit of analysis are known as single-subject designs.

## **What kind of research design is needed?**

In order to be able to distinguish among candidate experts, it is a good idea to present a wide range of stimuli. The researcher may not know how to identify stimuli that span the range; subject matter experts (SMEs) may be useful in selection of stimuli (although, in general, we do not wish to assume expertise, we acknowledge that SMEs do exist and can be helpful). The wider the range of stimuli, the more discrimination it is possible to exhibit. In some cases, expertise may show itself only when rare problems come along.

It is crucial that at least some stimuli be presented more than once. This repetition is necessary to provide an estimate of consistency. If it is not practical to present the entire set repeatedly, it is a good idea to select values from across the stimulus range to be presented repeatedly. There is a danger that amount of inconsistency depends upon stimulus magnitude (a la Weber's law),

If it is feasible, we recommend complete replication (presenting the entire set of stimuli more than once); the more replications, the more reliable is the estimate. Whether responses in the research setting can be sufficiently isolated to approximate independence is a standard concern for researchers.

It is imperative that the same set of stimuli be presented to all candidates. Varying the order of presentation of those stimuli, perhaps by employing independent randomization, across candidates may be acceptable, if order doesn't have an impact of its own.

### **What are the constraints on the stimuli to be evaluated?**

Stimuli need to be presented identically to all candidates. For ephemeral stimuli, that may present a practical problem. Video recording is a valuable tool, although some information may be lost relative to live presentation. For stimuli that are consumed during the task, sufficient quantities need to be on hand and they must not decay over time.

Stimuli also need to be presented more than once to an individual. If the stimuli are memorable, the candidate may try to recall rather than render an independent response. The researcher may need to space trials over time or re-label them so as to inhibit recollection.

### **What sorts of responses are usable?**

Because our approach to evaluating expertise is quantitative, we require discrete responses. These can be expressed on any of the response scales that experts use. With numerical ratings, the responses are approximately on an interval scale and can be used to construct variance ratios. Ordinal responses, such as rankings or letter grades, can be used similarly when converted to numbers. Sufficiently dense ordinal data yields considerable power (Weiss, 1986). Even “Yes-No” responses can yield results essentially equivalent to those obtained with continuous scales (Lunney, 1970). Nominal responses generally provide less power; nevertheless, measures of discrimination and inconsistency have been defined.

## What statistical procedures are applicable?

When CWS estimates of discrimination and inconsistency are variances, there is a statistical comparison that provides significance statements. Schumann and Bradley (1959) developed a procedure for comparing sets of data from two similar experiments that can determine whether one F-ratio is significantly larger than the other. Similar means that each F-ratio is constructed from responses to the same stimuli and therefore, has the same degrees of freedom. The ratio of the F-ratios constitutes a test statistic,  $\underline{w}$ .  $\underline{w}$  is compared to  $\underline{w}_0$ , a critical value found in the table presented by Schumann and Bradley. The test can be employed either directionally or nondirectionally. The one-tailed (directional) test determines whether the candidate is significantly less capable than a designated expert. The two-tailed (nondirectional) test asks whether there is a significant difference between two judges. Each judge is considered as a separate “experiment”. A computer program (Weiss, 1985) incorporating the Schumann and Bradley procedure and table of critical values is available.

The obtained  $\underline{w}$ 's represent the final steps in the CWS analysis. They allow comparison of the expertise exhibited by the various candidates as they judge a particular set of stimulus objects. Pairwise comparisons express how each candidate does with respect to the others. Alternatively, one may compare the candidate's expertise to an established standard.

### How do you average CWS scores?

Because CWS is built upon the squared differences between the impact of one stimulus and another, you must first get to a distance metric by taking the square root, then compute the mean of those roots and square the result.

$$\overline{\text{CWS}} = \left( \sum \text{CWS}_i \right)^2$$

For the four doctors in the example below,  $\text{CWS}_{18} = 580.20$ .  $\text{CWS}_8 = 1.21$ ;  $\text{CWS}_{16} = 1.81$ ;  $\text{CWS}_{23} = .76$

Mean CWS = 46.96 (the arithmetic mean is 145.99).

But please be careful; it is meaningful to average only when the stimuli were the same for all of the people whose individual CWSs are being integrated. A typical appropriate set of results to average would be the CWSs for a group of subjects at an early stage of training and then again at a later stage of training. Also be wary that because CWS values depend on the stimulus set, it would not be meaningful to compare the two means unless the stimuli were the same at both stages.

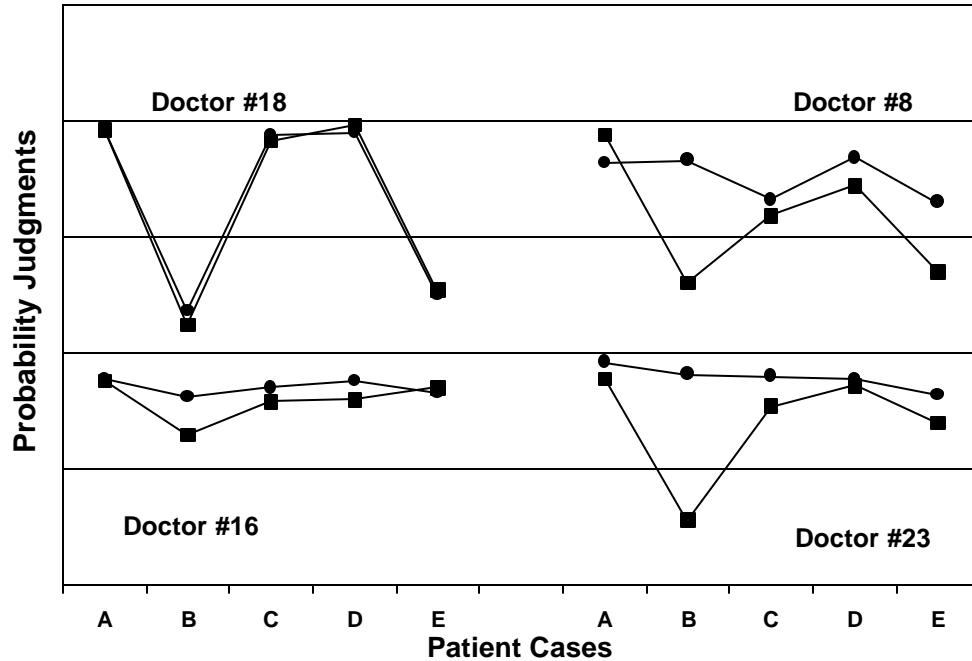
## **An example**

A recent study by Skånér, Strender, & Bring (1998) illustrates how expertise can be evaluated based on a set of judgments. Twenty-seven Swedish General Practitioners (GPs) judged the probability of heart failure for 45 cases based on real patients; five of the cases were repeated, although the GPs were not informed of that. The case vignettes stated that each patient came to the clinic because of fatigue. There were no additional pathological findings based on further examination. Normal values were provided for hemoglobin, electrolytes, s-creatinine, and TSH. Case-specific information was provided for ten cues: age, gender, history of myocardial infarction, dyspnea, edema, lung sounds, cardiac rhythm, heart rate at rest, heart X-ray, and lung X-ray.

“For each vignette, the doctors were asked to assess the probability that the patient suffered from any degree of heart failure” (Skånér et al., 1998, p. 96). The assessments were made on a graphic scale with “totally unlikely” at one end and “certain” at the other; these were converted into 0-to-100 values. The doctors were instructed “to judge the probability, not the severity, of heart failure” (p. 96).

Selected results for four of the GP’s (identified by number) are shown below. The five repeated cases are represented by letters at the horizontal axis. The circles are the judgments for the first presentation and the squares are the judgments for the second presentation. Thus, the first judgment of Case A by Doctor #18 is near 100; the second judgment is similar.

## Probability Judgments of Heart Failure



As can be seen, there is considerable variation between and within the four GP's. Still, each GP shows a distinctive pattern in terms of discrimination and reliability. Doctor #18 is highly discriminating (sizable differences between patients) and consistent (little difference between first and second presentations). Doctor #8 shows some discrimination, but lacks consistency (especially for patient B). Doctor #16 is consistent, but treats all patients rather similarly – all are seen as having moderately high chances of heart failure. Doctor #23 shows no uniform pattern of discrimination or consistency.

Based on their data alone, we can gain considerable insight into the judgment strategies and abilities of the GPs. Doctors #18 and #23 are consistent, but one discriminates and the other does not. Doctors #8 and #23 are inconsistent and vary in their discriminations. We believe that without knowing anything further, most clients would prefer someone like Doctor #18, who can make clear distinctions in a consistent way. The numbers shown in the graph were processed to yield the CWS values below. These CWS values are F-ratios.

Dr. #18	Dr. #8	Dr. # 16	Dr. #23
$CWS = 3365.15/5.80$ $= 580.20$	$CWS = 490.75/404.60$ $= 1.21$	$CWS = 65.40/36.10$ $= 1.81$	$CWS = 330.40/434.00$ $= .76$

## References

- Cochran, W. G. (1943). The comparison of different scales of measurement for experimental results. Annals of Mathematical Statistics, 14, 205-216.
- Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. Journal of Educational Measurement, 7, 263-269.
- 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.
- 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. (1985). SCHUBRAD: The comparison of the sensitivities of similar experiments. Behavior Research Methods, Instrumentation, and Computers, 17, 572.
- Weiss, D. J. (1986). The discriminating power of ordinal data. Journal of Social Behavior and Personality, 1, 381-389.