Judgment and Decision Making, Vol. 4, No. 2, March 2009, pp. 164–174

Criteria for performance evaluation

David J. Weiss*[1], Kristin Brennan[1], Rick Thomas[2], Alex Kirlik[3], and Sarah M. Miller[3]

[1] California State University, Los Angeles

[2] University of Oklahoma

[3] University of Illinois

Abstract

Using a cognitive task (mental calculation) and a perceptual-motor task (stylized golf putting), we examined differential proficiency using the CWS index and several other quantitative measures of performance. The CWS index (Weiss & Shanteau, 2003) is a coherence criterion that looks only at internal properties of the data without incorporating an external standard. In Experiment 1, college students (n = 20) carried out 2- and 3-digit addition and multiplication problems under time pressure. In Experiment 2, experienced golfers (n = 12), also college students, putted toward a target from nine different locations. Within each experiment, we analyzed the same responses using different methods. For the arithmetic tasks, accuracy information (mean absolute deviation from the correct answer, MAD) using a coherence criterion was available; for golf, accuracy information using a correspondence criterion (mean deviation from the target, also MAD) was available. We ranked the performances of the participants according to each measure, then compared the orders using Spearman's r_s. For mental calculation, the CWS order correlated moderately (r_s =.46) with that of MAD. However, a different coherence criterion, degree of model fit, did not correlate with either CWS or accuracy. For putting, the ranking generated by CWS correlated .68 with that generated by MAD. Consensual answers were also available for both experiments, and the rankings they generated correlated highly with those of MAD. The coherence vs. correspondence distinction did not map well onto criteria for performance evaluation.

Keywords: arithmetic, CWS index, judgment, measurement.

1 Introduction

To evaluate the work of a plumber, therapist, or surgeon, it is necessary to assess on-the-job performance. While all professionals have their creative moments, in most fields it is the ability to perform a practiced task consistently well that is the hallmark of the expert. Performance assessment is also the key to determining whether a training program or technical innovation is worthwhile. Ideally, assessment can be objective rather than a matter of opinion. Quantitative assessment of performance attends to measurable aspects of the work, typically the "bottom line" of the outcome of the labor. How many leaks were stopped? How many patients were cured?

Such outcome measures capture what Hammond (1996) refers to as correspondence competence, in that they focus directly on consequences. Outcomes can also be compared to theory-based standards; for example, updating of opinions should be governed by Bayes's theorem. Hammond (1996) refers to this type of standard as a coherence criterion. These two types of criteria for optimality compare performance to a gold standard, a compelling benchmark against which to measure the behavior. Indeed, some researchers argue that performance can be measured meaningfully only when a gold standard has been agreed upon (Ericsson, 1996). Just as Hammond (1996) hoped that the correspondence-coherence distinction would help to clarify debates about the proper way to evaluate a scientific theory, in this paper we invoke that distinction in the hope of clarifying debates about how to assess performance.

For many professional domains, gold standards simply are not available. What is the outcome that reflects the quality of a film review, the grade assigned by an instructor, or the sentence imposed by a magistrate? Weiss and Shanteau (2003) responded to the challenge that gold standards are elusive by constructing an empirical index, referred to as CWS,[1] that does not incorporate ground truth. They suggested that proficiency has evaluative skill

*Preparation of this manuscript was supported by the U. S. Air Force Office of Scientific Research (grant FA9550–04–1-0230 to California State University, Los Angeles) and by NSF (grant DRMS-045216 to the University of Illinois). We wish to thank Arash M. Taefi for writing the program used to present the intuitive arithmetic tasks. Correspondence regarding this article, including requests for reprints, should be sent to David J. Weiss, 609 Colonial Circle, Fullerton CA. 92835 United States. Email: dweiss@calstatela.edu.

[1]The CWS acronym derives from the initials of its creators, David J. Weiss and James Shanteau, along with that of the statistician William Cochran, who independently had previously proposed using an F-ratio to compare response instruments.

at its core. Whatever the task, one must attend to relevant aspects of the situation and decide what to do. Viewing evaluation as akin to what a measuring instrument does, Weiss and Shanteau (2003) identified two necessary properties of expert judgment: discrimination, responding differently to different stimuli, and consistency, responding similarly to similar stimuli. The CWS index, presented as Equation 1, combines these two properties in a ratio format. The ratio is large when the judge discriminates effectively, and is reduced when the judge is inconsistent. Weiss and Shanteau stressed that the two properties are not conceptually independent. It is easy enough to adopt a strategy that trades off one property at the expense of the other, but achieving both at the same time requires accurate evaluation of the stimuli, the essence of expert judgment.

$$CWS = \frac{Discrimination}{Inconsistency} \quad (1)$$

When they originally proposed the CWS index, Weiss and Shanteau were intentionally non-committal about the measures of discrimination and inconsistency. The trade-off implied by the ratio definition is the heart of the concept, and any measures that reflect the two properties will do. In applications that generate numerical data, including the present ones, discrimination and inconsistency have been operationalized using terms familiar from analysis of variance. An experimental design suitable for CWS analysis may be as simple as the presentation of each of several stimuli more than once. Discrimination means that different stimuli are responded to differently. Accordingly, discrimination is captured by the mean square between stimuli. Inconsistency implies that a given stimulus presented multiple times inspires different responses on the various occasions. Inconsistency is captured by the mean square between replications.

The CWS approach resembles a coherence criterion, in that it examines purely internal properties of behavior. However, it differs from other coherence criteria in that while proficient performance inexorably generates high values of CWS, there is no theory specifying the optimal behavior. Our view is that performance ought to be tied to the external world, and that experts should follow the prescriptive model for their task. However, it is not always possible for an evaluator to know the best answers, and the applicable model is often unknown as well. The absence of optimal answers does not diminish the practical importance of having the capability to evaluate members of the large class of professionals who provide opinions about the status and achievements of people (Weiss, Shanteau, & Harries, 2006).

A more popular approach to evaluating these subjective domains is to compare someone's responses to those of other people. Opinions often converge toward the truth (Surowiecki, 2004). Consensual answers have often been proposed as surrogates for correct answers (Ashton, 1985; Einhorn, 1974), although the logic of doing so has been criticized (Weiss & Shanteau, 2004). The gist of the criticism is simply that people may agree on poor answers. One may view consensus as a coherence criterion, postulating that there exists across people a common latent structure underlying their opinions (Batchelder & Romney, 1988; Uebersax & Grove, 1990).

In the current project, we employed tasks for which there were indisputably optimal responses, namely mental calculation and golf putting. Accuracy in arithmetic calculation is customarily assessed using a coherence criterion; correct answers are dictated by the abstract, logical rules of mathematics. The accuracy of a putt is usually assessed using a correspondence criterion, how close the ball gets to its target. A goal of the present research was to shed light on CWS's ability to capture the subjective domains by examining objective domains. We assessed performance for both tasks using the clear-cut gold standards, then assessed that same performance using CWS, which does not make use of such information, and consensus, which provides a "silver standard" (Phillips, 1988) when the group knows what it is doing.

1.1 The logic underlying CWS

A CWS assessment entails analyzing responses to a range of stimuli, situations, or scenarios that would normally be handled within the profession. Tasks may be divided into four categories (Weiss & Shanteau, 2003). *Judgment* is exemplified by auditing a financial statement or diagnosing a patient's condition. *Prediction* includes forecasting the weather or advising the parole board. *Teaching* encompasses training people or setting criteria for testing. Typical physical *performance* tasks are playing an instrument or shooting a ball. In all cases, evaluating the stimuli underlies proper execution of the tasks. In the latter three categories, additional abilities overlay the requisite judgmental skill. The predictor must anticipate changes that will occur in the future. The instructor must communicate and motivate. The performer requires physical abilities needed to execute the planned maneuvers. The CWS index can be used to assess behavior in all of these categories, but underlying judgmental skill may be obscured by the additional components.

Still, because judgment is paramount, reasonably accurate assessments of demonstrated skill in all of the categories can be achieved with CWS. The key properties, discrimination and consistency, are inherent in the behavior itself, so that measuring the ratio does not require knowledge about how things turned out. Of course, there is more to skilled performance than these two properties. CWS is necessary but not sufficient; in other words, one

https://doi.org/10.1017/S1930297500002606 Published online by Cambridge University Press

who does the task well will generate high CWS, but high CWS does not guarantee that the task was done correctly (Weiss & Shanteau, 2003). The question of how much of the demonstrated skill is captured by the index is essentially an issue of validity.

In order to assess validity, one must have some approximation of the truth. We suggest five presumptions an analyst might make toward that end. Each presumption assumes domain knowledge on the part of the analyst, external knowledge that is provided by experts within the field. This circular reasoning, presuming that the analyst can identify the true domain experts, seems unavoidable in the early stages of research. The first three of the presumptions have been supported in previous research using the CWS framework. The last two have not been tested before.

Presumption 1 is that CWS can distinguish experts from novices; experts should generate higher CWS scores. Weiss & Shanteau (2003) illustrated this capability with data from several domains, including medicine, auditing, and personnel selection. Identifying novices is easy, but we have to assume that we know who the experts are in order to validate. Regarding experience as the equivalent of expertise is risky (Weiss, Shanteau, & Harries, 2006).

Presumption 2 is that CWS decreases systematically with increasing task difficulty. Here, the assumption is that the analyst can identify the more difficult tasks. Shanteau, Friel, Thomas, and Raacke (2005) varied the number of planes in simulated air traffic control, reasoning that having to deal with more aircraft should make the task harder. CWS decreased with the number of planes.

Presumption 3 is that CWS increases with training. Shanteau et al. (2005) also found that CWS increased over training periods, showing improvement long after less sensitive (outcome) measures such as the number of accidents or number of intrusion errors stopped showing performance gains. The assumption in this case is that the analyst knows performance to be in the sub-asymptotic range where increases are possible. In a study of the assessment of upper limb disorders, Williams, Haslam, and Weiss (2008) found that professional ergonomists, who had specialized training, exhibited higher CWS when judging patients' risk status than members of other professions who also make such judgments regularly. The ergonomists also were superior, according to CWS, to students in ergonomics courses. Similar results for occupational therapy have been reported by Rassafiani, Ziviani, Rodger, and Dalgleish (2008).

The two new experiments reported here examine our fourth and fifth presumptions. The experiments are quite different in nature, but they have in common that there are known correct answers. In the first experiment, college students are asked to carry out intuitive addition and multiplication under time pressure. The tasks in Experiment 1 are purely judgmental. In Experiment 2, golfers putt toward a series of targets. This task involves physical performance as well as an implicit judgment. Our purpose in selecting both a cognitive and a perceptual-motor task was to shed light on the breadth of applicability of CWS as a performance index.

Presumption 4 is that CWS should be associated with the extent to which performance follows a correct process model. For the mental calculation tasks, participants who show higher CWS should be more likely to follow the additive and multiplicative models as assessed by functional measurement. Functional measurement (Anderson, 1979) invokes a coherence criterion, in that there is a normative model for the task; the analysis involves examining the algebraic structure underlying the responses. Because the number of everyday tasks for which a prescriptive model is available is limited, this presumption can be examined only in special cases. Presumption 4 cannot be tested with the putting task.

Presumption 5 is more widely applicable. Presumption 5 is that, in tasks for which correct answers are available, CWS should be higher for those whose answers are closer to correct. Our analytic strategy for testing Presumption 5 is to first rank the performances exhibited by the individuals within an experiment according to the gold standard of correct answers. Next, we rank the same performances according to CWS, which knows nothing of correct answers. High correlation between the two rankings is supporting evidence for this presumption. This comparison is the key empirical contribution of the present paper. If CWS can be shown to capture differences in performance when correct answers are known, that increases confidence in the ability of this relatively new index to provide valid assessments when correct answers are unavailable.

In order for this research strategy to be effective, it is necessary that there be differential ability among the participants. At the same time, they must all be able to do the task with some degree of competency, or the results mean little. We were willing to presume that all college students can do mental calculations. For golf, we required credentials in the form of some experience, as novices have essentially no expertise. To some extent it is a matter of fortune whether the recruits in a study vary sufficiently, but we tried to assist chance by informally seeking a range of self-estimated talent for arithmetic.

Employing a gold standard of correctness requires the analyst to choose a rule for penalizing errors. When the response is measured on a numerical scale, it is traditional to use the mean squared deviation (MSD) from the correct answers as an index. Gigone and Hastie (1997) provide an extensive comparison of accuracy measures, favoring MSD because it contains the most information and penalizes large errors, which they see as an advantage. Our view is that while MSD fits nicely with statistical theory, it does not reflect how wrong the answer is from a behavioral perspective (Weiss, Edwards, & Shanteau, 2009).[2] Although we will report MSD, we deem the mean absolute deviation (MAD) from the correct answers to be the gold standard for performance.

The unique feature of this study is that we use the same data to compare various performance criteria. In evaluating Presumption 5, we use the golf data to provide a direct comparison of a correspondence criterion (accuracy) and a coherence criterion (CWS). Similarly, we invoke Presumption 4 with the mental calculation data to suggest a comparison of two kinds of coherence criteria, one (functional measurement) that incorporates a standard of optimality and one (CWS) that does not. We also evaluate the coherence criterion of consensus with both data sets.

2 Experiment 1: Mental calculation

The task for participants is to solve math problems in their heads (Busemeyer, 1991; Peterson & Beach, 1967); specifically, to perform mental calculation of either the sum or the product of a pair of numbers. Preliminary work using these tasks suggested that incorporating some time pressure was necessary in order to induce holistic judgments. Explicit use of arithmetic rules was deemed undesirable, because we wanted the laboratory task to simulate real-world judgments, few of which have formulaic solutions. The participant's incentive on each trial was based on the difference between the response and the correct answer; no credit was given for responses occurring after a time limit specified for each problem type.

Intuitive addition of numerical stimuli has been extensively studied, including research that employed a functional measurement perspective (Anderson, 1968; Levin, 1975). A result of particular interest is that some people exhibit consistent biases, thus implying incorrect answers, while following the appropriate model. This illustrates the key principle that a focus on accuracy may obscure important information. Multiplication is inherently more difficult than adding, and one would expect less accurate answers. Intuitive multiplication has not been studied much beyond one or two-digit problems (Seitz & Schumann-Hengsteler, 2000).

2.1 Method

Participants and Procedure. Twenty participants were recruited via fliers posted across the California State College, Los Angeles campus, with the qualification being that applicants were enrolled students (any major) and at least 18 years old. Students who claimed to be poor at math were encouraged to participate; those recruited spanned a wide range of (self-assessed) ability. Participants received base compensation at minimum wage level, as well as a bonus for accurate answers.[3] They also received a bonus for completing all sessions, which took between 2–3 hr.

After receiving instructions regarding use of the computer program, participants performed the arithmetic tasks on a computer in a small individual laboratory, with the experimenter visible in the hallway. The concept of intuitive math was stressed; the use of paper or calculator was prohibited. Participants were told to guess if they did not know the answer, as there was no penalty for wrong answers. They were informed that an answer within 5% of the correct value would be scored as correct for bonus purposes, but the answer had to be entered within 30 sec. There was an additional brief training component for each type of equation, with outcome feedback and illustrative bonus points. No feedback was provided during data collection.

The program presented two types of problems, one calling for adding and the other for multiplying. Most of the problems were expected to be difficult for most students. In an attempt to inhibit explicit calculation, each problem was presented briefly (3 sec for addition, 5 sec for multiplication) before the screen went blank.

The inter-question interval was 15 sec, but the participant could bypass the break by clicking the Enter key. After a block of trials, which lasted 10–20 min, the program informed the participants of how many bonus points had been earned during that block. Three to four blocks of trials were scheduled during each 1 hr session. In between blocks, participants were allowed to take brief rest periods.

Design. The anticipated difficulty of the problems was manipulated by varying the number of digits in the num-

[2] In everyday life, errors are often penalized on an absolute basis, and occasionally on the basis of extent. For example, a basketball shot either goes in or misses. In golf, the distance the ball lands from the hole contributes to the difficulty of the next shot. We are hard pressed to think of natural situations in which errors are punished in proportion to the square of their magnitude. Using simulation results, Dielman (1986) concluded that the use of absolute value in regression analysis provides better forecasts than does the use of least squares, especially when the data contain outliers.

[3] An answer within 5% of the correct value earned a bonus of $.05. Typical performance earned a bonus of $4–5/hr.

https://doi.org/10.1017/S1930297500002606 Published online by Cambridge University Press

Table 1: Numbers used for addition and multiplication problems within each difficulty level.

2 digit, 2 digit		2 digit, 3 digit		3 digit, 3 digit	
Left Position	Right Position	Left Position	Right Position	Left Position	Right Position
18	15	26	195	131	138
29	24	34	268	294	216
33	51	49	391	352	425
46	57	52	453	384	548
64	64	65	575	585	641
79	72	77	628	613	776
83	87	81	746	882	893
96	98	93	982	947	991

ber. Three difficulty levels were used. The same difficulty level was maintained throughout a block. The pairs of numbers that constituted the 64 problems within each difficulty level were constructed according to an 8x8 (left position x right position), factorial design, following Anderson (1968). No number was permitted to have a zero as the right-most digit. Addition problems came first, then multiplication; in both cases, the problems were presented in order of increasing difficulty level. Two replications of the design, using the same 64 pairs of numbers in an independently randomized order, were presented for both addition and multiplication problems, thus yielding a total of 12 blocks for each participant. The numbers used are shown in Table 1.

2.2 Measures

Accuracy. Although accuracy sounds transparent enough, there are at least three sensible ways to capture the accuracy of the responses. The most commonly used measure, *Mean Squared Deviation* (MSD), is the average of the squared deviations, $(\sum [C_i - X_i]^2)/N$ where C_i is the correct answer and X_i is the response on the i-th trial. The *Mean Absolute Deviation* (MAD), $(\sum |C_i - X_i|)/N$ is an accuracy measure that does not weight discrepancies via squaring. The *Correlation* between correct answers and responses is also an accuracy measure, but it does not distinguish between truly accurate and linearly discrepant responses (Stewart & Lusk, 1994). All of these accuracy measures may be viewed as coherence based, in that they compare correct answers to those specified by a mathematical formula.

CWS. The CWS index (Weiss & Shanteau, 2003) for an individual's performance is the ratio of discrimination to inconsistency, calculated separately for each task and difficulty level as the mean square between stimuli divided by the mean square within replications. The computation for Equation 1 is that for a single-S design (Weiss, 2006), which is identical to the calculation of an F-ratio in an independent groups design. Accordingly, the data can be entered into a standard ANOVA program as a 64 (stimuli) x 2 (replications) design. CWS may be viewed as a coherence standard, in that it is based on a theory of optimality, but of a special type that does not incorporate correct answers. The CWS ratio depends only upon internal properties of the set of responses.

Consensus. The mean response can be calculated across respondents for each stimulus pair and that mean can be used as a criterion. From the set of consensual criteria, we can construct *pseudo-accuracy* measures similar to the accuracy measures described above. Our consensus measure was based on MSD, in that we substituted the mean response, M_i, for the correct answer, so that *Consensus* is $(\sum [M_i - X_i]^2)/N$. We might equally have based the consensus measure on MAD, but did not because the MSD-based version has been traditionally used in the literature.

For an arithmetic task, those who provide correct answers will inevitably agree. However, agreed-upon answers need not be correct. One possibility is the widespread use of a heuristic strategy (Gigerenzer, Todd, & The ABC Research Group, 1999) that simplifies the challenging multiplication task. For example, one might round the numbers to the nearest ten or hundred prior to multiplying, and then make an upward or downward adjustment to correct for rounding.

Model Fit. Because the stimuli were constructed according to a factorial design, it is possible to employ functional measurement analyses (Weiss, 2006) on the responses. Functional measurement invokes a coherence criterion, evaluating the fit of a plausible algebraic model to the observed judgments. For the adding task, an additive model should apply. For the multiplying task, a multiplicative model should apply. These models can be tested using analysis of variance; they predict that specific sources in a factorial analysis of variance will yield significant effects and that others will not (Weiss, 2006). If people do the task perfectly, then the model will fit, and the answers will be accurate. However, it is possible for the model to fit and for the answers to be systematically inaccurate, e.g., by consistently placing higher weight on the number presented on the left (primacy). A potential weakness of the model fit approach is that high variability increases the likelihood that data will appear to support the model because there is insufficient power to reject it.

https://doi.org/10.1017/S1930297500002606 Published online by Cambridge University Press

Table 2: Performance for two individual participants across mental calculation tasks, as assessed by six indices.

	Problems					
Index	2 + 2	2 + 3	3 + 3	2 x 2	2 x 3	3 x 3
Participant G						
MAD	0.06	0.08	0.40	3.12	24.59	250.56
MSD	0.75	12.64	80.45	12396.00	323660.13	39112205.50
Correlation	0.94	0.98	0.96	0.78	0.94	0.93
CWS	16.84	41.05	22.63	3.16	18.94	15.56
Consensus	9.98	15149.85	1519.92	66886.54	6480550.31	436290435.69
Model Fit	1.03	1.04	0.82	1.12	0.92	1.08
Participant E						
MAD	0.90	17.49	0.68	3.78	625.25	2236.50
MSD	60.84	6996224.11	2135.85	14784.45	2886585990.68	68662171252.01
Correlation	0.54	0.30	0.64	0.68	-0.12	0.22
CWS	1.24	1.00	1.66	2.44	1.15	1.00
Consensus	78.69	9017726.92	3325.01	65329.33	2870853841.73	67531546588.76
Model Fit	1.00	1.00	1.08	1.06	0.62	0.13

2.3 Results

Performance. To convey the flavor of the data, we present the performance indices achieved by one of the most successful participants and by one of the least successful in Table 2. These extremes illustrate how the various indices track the same observed behavior. The three accuracy measures (MAD, MSD, Correlation) all report the superiority of Participant G over Participant E in the same way, with lower values for all six of the problem types. MAD and MSD also confirm that multiplication is more difficult than addition and that problem difficulty increases with the number of digits, although Participant E had an especially hard time with adding 2 digit and 3 digit numbers. On the other hand, Correlation was not effective in capturing these expected trends.

Our featured index, CWS, did show the superiority of Participant G's performance over that of Participant E, but did not fare particularly well in capturing the difficulty we built into the design. The picture presented when Consensus was used as a surrogate for correctness was comparable to that provided by MAD and MSD.

We assessed model fit using the F-ratio of the source that captures deviations from the normative additive (testing the Left x Right interaction) and multiplicative (testing the deviations from bilinearity) models (Weiss, 2006) for the respective tasks. These F-ratios are shown in the bottom line of Table 2. The normative models were quite descriptive, in that the key F-ratios were nonsignificant for most blocks for most participants. This nonsignificance is not attributable to lack of power, because the main effects and (for the multiplicative model) the bilinear component of the interaction were both significant and sizable. Graphically, the appropriate pattern — parallelism in the case of addition, fan in the case of multiplication — was observed for most of the individual plots, especially for addition. Thus, the functional measurement analysis does shed light on the behavior, telling us that people did follow the applicable combination rule. However, the small F-ratios did not distinguish among participants of varying proficiency. In this application, the process analysis was uninformative regarding differential proficiency; Presumption 4 could not be verified. To be fair, functional measurement has never been proposed by its adherents as a tool for assessing performance; nor has the magnitude of a nonsignificant F-ratio ever been proposed to be meaningful.

Participant Rankings. One of our primary purposes was to see how the indices compared in terms of scoring the people. In employment contexts, rankings are the usual basis of decisions. For each index, we ranked the 20 participants according to their average score[4] across the

[4]To average the index values across the six conditions for each participant, we followed the recommendation of Weiss and Edwards (2005), transforming so that the averaging is carried out on the units of the original measurements. For CWS, FM, MSD and Consensus, the appropriate average is the square of the mean of the square roots of the six individual values. For Correlation, we employed Fisher's r to z

https://doi.org/10.1017/S1930297500002606 Published online by Cambridge University Press

Table 3: Rank order correlations (Spearman's r_s) between six performance measures on mental calculation tasks.

	MSD	Correlation	Model Fit	CWS	Consensus
MAD	.78*	(–).65*	(–).05	(–).46*	.70*
MSD		(–).55*	(–).34	(–).46*	.92*
Correlation			.07	.86*	(–).50*
Model Fit				.37	(–).55*
CWS					(–).43

* $p < .05$. n = 20 for all correlations. Minus signs indicate direction only, and are unimportant to the strength of the relationship.

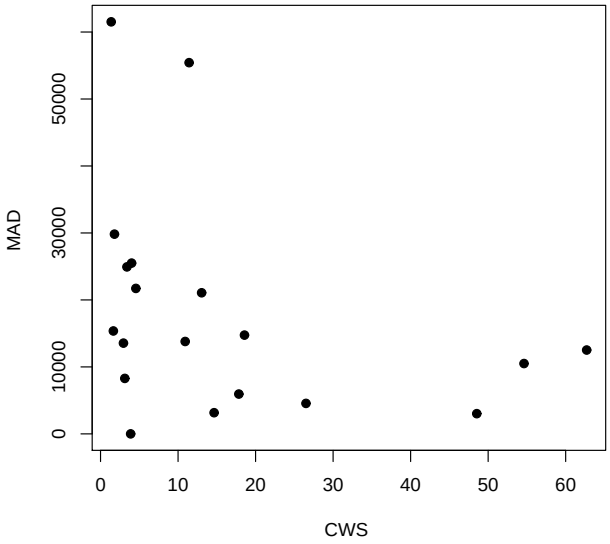

Figure 1: CWS vs. MAD for mental calculation data from nineteen students. Each data point represents the appropriate index-specific average over the six conditions. Spearman's r_s = (–).50, p <.05. In order to avoid distorting the graphical impression of the relationship, we omitted the data from an outlier whose average CWS was much higher than anyone else's. With that twentieth student included, r_s = (–).46, p <.05.

6 conditions, then examined the correspondence among those rankings using Spearman's r_s, the rank-order correlation. The rank orders we compared were based on quality of performance as conveyed by each measure. For MSD, MAD, Consensus, and Model Fit, lower scores indicate better performance. For Correlation and CWS, higher scores indicate better performance. These correlations are presented in Table 3.

We consider MAD to be the gold standard for the task. The other accuracy indices yielded rankings that agreed well, but not perfectly, with that established using MAD. Presumably, MSD generates slightly different orders because it weights large errors differently. Correlation is a less sensitive index, in that it can fail to penalize responses that are incorrect if the errors follow an orderly pattern. Consensus did fairly well, perhaps reflecting the objective nature of the task. People are aiming at the same target, the correct answer, and on average their guess corresponds to that target reasonably well (Surowiecki, 2004). Consensus and MSD, which both square deviations, yielded similar rankings. CWS agreed moderately with the gold standard order. CWS agreed moderately with the gold standard order as is shown graphically in Figure 1.

transformation. The ordinary arithmetic mean is appropriate for MAD.

3 Experiment 2: Golf putting

There is an obvious gold standard for a golf shot; the ball either goes in the hole or does not. Within the traditional game, the degree of imperfection of a shot that misses is measured by the number of subsequent shots required to get the ball in. However, the latter measure is confounded with the quality of those subsequent shots. A more pure measure of the imperfection of a shot is the distance between its landing point and the hole.

We employed a laboratory version of golf putting that has proven useful in understanding skilled performance and its attentional limitations (Beilock & Carr, 2001; Beilock, Carr, MacMahon, & Starkes, 2002; Perkins-Ceccato, Passmore, & Lee, 2003). In this stylized abstraction of one of golf's core skills, the task is to putt to a target. The distance between where the ball lands and the target is analogous to the difference between the correct answer and the stated response in our arithmetic tasks.

For golf, we can invoke a correspondence assessment. The ball is supposed to hit the target, and experience teaches golfers how to achieve that goal. We can also parse each golfer's putts into a CWS index. In accord with Presumption 5, we anticipated that higher CWS

https://doi.org/10.1017/S1930297500002606 Published online by Cambridge University Press

would be related to more accurate putting. Our logic is that greater discrimination means that the golfer knows to hit the ball farther the more distant the target. Better golfers should also be more consistent, because their strokes are well-regulated. A Consensus criterion, again invoking the argument about a common latent structure guiding the golfers' efforts, is also available.

3.1 Method

Participants. We report data from twelve experienced golfers between the ages of 18–22. Participants were required to have two or more years of high school varsity golf experience or a Professional Golfers' Association (PGA) handicap less than 8. The session lasted approximately one hour. The golfers were paid $10 for their time. There was no performance-based incentive.

Experimental design and procedure. Participants received instructions to putt the ball from one of nine different starting points so it stopped as close as possible to a target, marked by a taped X on a uniformly flat synthetic turf mat. Three of the locations were 1.2 m from the target, three were 1.4 m from the target, and the other three were 1.5 m from the target. Following instructions and 10 practice putts, participants performed two blocks of 21 putts. Accuracy (more precisely, amount of inaccuracy) for a single shot was measured by the distance from the center of the ball to the center of the target (in cm). This target feature is slightly more challenging than real golf, in that there is no hole in which the ball can come to rest. Possibly, a shot that rolled gently over the target (and thereby generated an error) might have gone into a real golf hole. The Mean Absolute Deviation (MAD) for an individual at each starting location was computed by averaging the single shot accuracy scores.

We also measured the total distance the ball traveled (in cm), for use in the CWS computation. CWS is defined as the ratio of discrimination to inconsistency. Discrimination, the numerator of the ratio, is calculated here as the mean square between the distances the ball was hit from different starting points. Inconsistency, the denominator, is divided by the mean square between replications, that is, the mean square between the distances when the ball was hit from the same starting point. Thus as in Experiment 1, CWS is computed like a standard F-ratio.

3.2 Results

The task was fairly challenging. Only 4 putts actually landed on the target. There were 22 additional putts that had a zero angle error and a distance greater than the correct distance; some of these might have gone in a real hole. So all told, about 10% of the putts could have gone in.

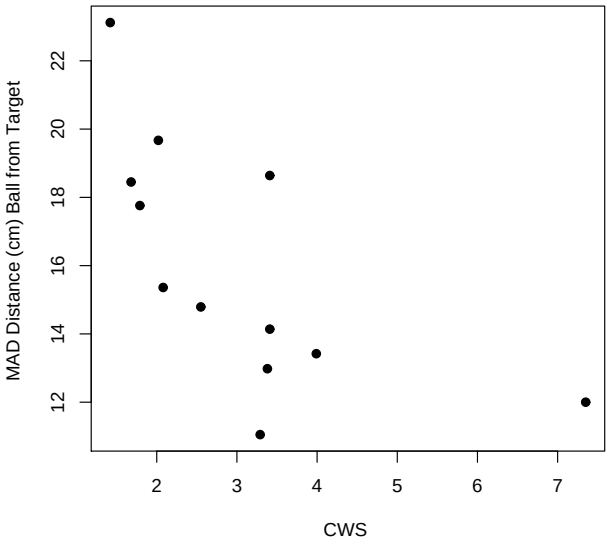

Figure 2: CWS vs. MAD for putting data from twelve golfers. Spearman's r_s = (–).676, $p < .05$.

We calculated a separate CWS value for each golfer, entering a 21 (stimuli) x 2 (replications) single-S design into the ANOVA program. The rankings from CWS were significantly correlated with those from accuracy as measured by MAD (r_s = (–).676, $p < .05$; n = 12). Thus, although CWS is ignorant of how far away the target is, or whether the shot is accurately directed, it yielded values that were reasonably well correlated with putting performance as measured by a gold standard (distance between the final ball location and the target). The relationship is shown graphically in Figure 2.

In the evaluation of Consensus, we used as the consensual answers[5] the mean distance the ball was hit toward each of the targets, and calculated each individual's deviations using the same definition employed for consensus in the mental calculation tasks, $(\sum [M_i - X_i]^2)/N$. The rankings for putting generated by the consensus criterion correlated very highly with the accuracy rankings given by MAD, r_s = .888 ($p < .05$, n = 12).

4 General Discussion

In the present paper, we illustrate how CWS can be used to assess complex performance in the absence of a true gold standard, using mental calculation and golf putting as prototypes. We chose these tasks because in each case there is an obvious gold standard against which to test the capability of the index. How much proficiency can be

[5] An alternative definition of the consensual answer might be the centroid of the landing points of the putts toward each target. Our data were not collected in a manner that would permit that centroid to be calculated.

https://doi.org/10.1017/S1930297500002606 Published online by Cambridge University Press

captured if we don't know the right answers or whether the ball lands on the target? The answer seems to be that a moderate amount of the proficiency can indeed be detected by CWS, a measure that looks only at the discrimination and consistency exhibited by the respondents. Combining these two necessary properties of good judgment yields an index that is able to capture a considerable amount of the variation in how well people did the tasks. The participants in these studies knew nothing of CWS. They were trying to maximizing accuracy, not discrimination and consistency. Because accuracy subsumes discrimination and consistency, CWS can serve as a proxy.

CWS is a coherence criterion, albeit an unusual one in that there is no gold standard for a response. Experiment 1 examined how CWS compared with other coherence-based measures. Our Presumption 4, that Model Fit would be associated with CWS, was not supported. More generally, the several criteria we employed for mental calculation did not yield correlated rankings. So we may conclude that not all coherence criteria produce similar evaluations. Experiment 2 compared CWS to a correspondence-based measure. Our Presumption 5, that participants who score well according to an accuracy criterion also should score well according to CWS, was confirmed by the correlation in the rankings for both mental calculation and golf putting. The golf results showed that a theory-based coherence criterion can produce evaluations similar to those generated by a correspondence criterion. That fact that evaluations using coherence criteria do not group themselves conveniently, and that evaluations using coherence criteria do not stand apart from that produced by a correspondence criterion, casts doubt on the value of Hammond's distinction in this context.

The technique we relied upon, comparing the rankings generated by the various indices, is constrained by the true differential expertise of the participants. The more similarly the contenders performed, the less differential performance there is for CWS (or any performance index) to detect. To that end, we selected tasks our participants already knew how to do but at which they were not so skilled that all performances would be excellent. Accordingly, the exact magnitude of the correlations between CWS and MAD is not critical to our validation; what does matter is that CWS has been shown to detect differences in demonstrated skill in much the same way that MAD did for both a judgment task and a physical performance task.

The CWS index has previously been used to assess the judgmental performance of professionals in several domains for which a true standard is unavailable. These include physicians judging the likelihood that patients had chronic heart failure (Weiss & Shanteau, 2003), occupational therapists prioritizing clients for therapy (Weiss, Shanteau, & Harries, 2006), and ergonomists determining the risk of workers complaining about upper limb disorders (Williams, Haslam, & Weiss, 2008). A lingering concern has been that if a putative expert consistently discriminates an irrelevant feature of the behavior, CWS can be fooled (Weiss & Shanteau, 2003). It is crucial to tap into a dependent variable that captures the heart of performance on the task. Identifying the right variable requires domain knowledge. For example, in CWS investigations of air traffic control (Shanteau et al., 2005), time through sector eventually came to be recognized as the dependent variable of choice. CWS indices built on time through sector were found to be related to task difficulty and to type of training. The evidence from the mental calculation and golf studies presented here suggests that CWS does indeed provide reasonable assessments of proficiency. So long as a sensitive independent variable has been chosen, performance assessment can proceed with the analyst blind to any individual characteristics of the contenders.

CWS's ability to capture performance on the putting task without target information is particularly impressive. Not only is the measure unaware of whether the ball was struck accurately, it does not even know which of the targets are farther away. Because the distance the ball traveled turned out to contain useful information, in principle it would have been possible to assess differential performance just by knowing how hard the ball was struck.

Our outcome-based assessments would not have been feasible without continuous error measures. Few of the answers in the arithmetic tasks (other than for two digit + two digit addition) were exactly correct. Capturing performance by whether a ball either goes in the hole or not can be insensitive to how well a shot was hit. We noted that only about 10% of the putts could have gone in a hole.

We advocate MAD as the error penalization rule; but our results suggest that the traditional MSD index, in which errors are squared, yields similar rankings. Because we calculated CWS ratios using mean squares, CWS effectively uses squared error penalization. The use of mean squares is not critical to the formulation of CWS, and indeed it is feasible to construct an index using measures of discrimination and inconsistency based on MAD. Whether the analyst's decision regarding error penalization will have serious consequences depends upon whether large errors occur frequently within the data set.

The CWS methodology for assessing performance is limited to quantifiable, repeatable behaviors. CWS is, like all objective assessment techniques, limited in scope. It does not address the quality of an actor's performance, an artist's creation, or a professor's lecture. As well, when comparing performers, every candidate must face essentially the same conditions. Repeatability within a

https://doi.org/10.1017/S1930297500002606 Published online by Cambridge University Press

person can also be a limitation; some tasks can be done meaningfully only once. The analyst applying CWS must be willing to assume that observations occurring at different moments are in fact comparable (Weiss & Shanteau, 2003). Under those circumstances, which characterize much of the routine work of many professionals, when correspondence measures are unavailable, a reasonable assessment of performance can be achieved with CWS, a coherence criterion.

5 Summary and confession

Hammond (1996) used the coherence-correspondence distinction to help distinguish among metatheories for scientific truth, where the metatheory provides a basis for telling whether a theory is true. We attempted to map that distinction onto performance evaluation. We had thought that comparing observed to optimal responses was employing a correspondence criterion, while the use of CWS, an index based on a theory of expert performance, was an application of a coherence criterion. Accordingly, we examined evaluation methods that did or did not incorporate optimal responses. The reviewers showed us the error in our thinking, in that arithmetic is at its core a theory-driven system that does not depend upon a connection with consequences. So involving optimality did not imply the use of a correspondence approach.

Our imperfect mapping leads us to suggest that the coherence-correspondence distinction may not be well-suited to dichotomizing performance criteria. Empirically, the distinction did not provide two distinct sorts of results. One reason that correspondence does not stand alone may be that correspondence criteria are likely to be temporary in applied contexts. For example, a physician's competence or a drug's value might at first be evaluated according to patient outcomes such as survival, a correspondence criterion. But as medical science progresses and theoretical insights evolve, the rather crude index of survival is replaced by physiological indicators. Of course the physiological measures are ultimately connected to survival; but they are connected by a theory, and it is that theory that governs the construction of the instrumentation that reports the measures. Lewin's (1951) famous dictum that "there is nothing more practical than a good theory" is very pertinent to performance evaluation.

An alternative dichotomization that might be proposed for evaluating performance is using process criteria vs. using outcome criteria. For example, one might evaluate an athlete's performance according to either purity of style or to scoreboard result. Purity of style is the basis in figure skating and diving. Most other sports, in contrast, are scored according to the final result: who was the fastest, who scored more points, etc. One might expect good form to produce good results, but it remains an empirical question whether the quarterback who throws the most beautiful spiral also completes the most passes. It is interesting to note that modern baseball analysts are beginning to evaluate players according to process — for example, how many pitches does a hitter swing at — as opposed to traditional criteria such as batting average (Lewis, 2003).

Applying objective process criteria to behaviors that take place out of sight, most prominently thinking, is challenging. Our Presumption 4 did invoke a process model for mental calculation. Although we had prescriptive algebraic models for the two arithmetic tasks, the measure of discrepancy from model predictions was not associated with the gold standard of correct answers or with most of the other indices. We concluded that functional measurement was not an effective tool for comparing candidates. The moderate correlation between Model Fit and Consensus is an anomaly within this description. We found it puzzling that Model Fit could be correlated with consensus, Consensus with everything else, and yet Model Fit with nothing else. Error variance may play a role in this seeming paradox; but a more satisfying resolution is that contrary to intuition, correlations are not necessarily transitive (Langford, Schwertman, & Owens, 2001).

In a more positive vein, we were able to confirm that the CWS index was effective in capturing performance, and we now have further justification for recommending it when valid outcome measures are not available. We also found that consensual answers can provide an effective substitute for correct answers when the task is one that people do reasonably well.

References

Anderson, N. H. (1968). Averaging of space and number stimuli with simultaneous presentation. *Journal of Experimental Psychology*, *77*, 383–392.

Anderson, N. H. (1979). Algebraic rules in psychological measurement. *American Scientist, 67*, 555–563.

Ashton, A. H. (1985). Does consensus imply accuracy in accounting studies of decision making? *Accounting Review, 60*, 173–185.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, *53*, 71–92.

Beilock, S. L., & Carr, T. H. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology: General, 130*, 701–725.

Beilock, S. L., Carr, T. H., MacMahon, C., & Starkes, J. L. (2002). When paying attention becomes counterproductive: Impact of divided versus skill-focused at-

https://doi.org/10.1017/S1930297500002606 Published online by Cambridge University Press

tention on novice and experienced performance of sensorimotor skills. *Journal of Experimental Psychology: Applied, 8*, 6–16.

Busemeyer, J. R. (1991). Intuitive statistical estimation. In N. H. Anderson (Ed.), Contributions to information integration theory, Vol. 1 (pp. 187–215). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dielman, T. E. (1986). A comparison of forecasts from least absolute value and least squares regression. *Journal of Forecasting*, *5*, 189–195.

Einhorn, H. J. (1974). Expert judgement: Some necessary conditions and an example. *Journal of Applied Psychology, 59*, 562–571.

Ericsson, K. A. (1996). The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson (Ed.) *The road to excellence: The acquisition of expert performance in the arts and sciences, sports, and games* (pp. 1–50). Mahwah, NJ: Erlbaum.

Gigerenzer, G., Todd, P., & The ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford, UK: Oxford University Press.

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin, 121*, 149–167.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York: Oxford University Press.

Langford, E., Schwertman, N., & Owens, M. (2001). Is the property of being positively correlated transitive? *American Statistician, 55*, 322–325.

Levin, I. P. (1975). Information integration in numerical judgments and decision processes. *Journal of Experimental Psychology: General*, *104*, 39–53.

Lewin, K. (1951). *Field theory in social science: Selected theoretical papers.* New York: Harper & Row.

Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. New York: W. W. Norton.

Perkins-Ceccato, N., Passmore, S. R., & Lee, T. D. (2003). Effects of focus of attention depend on golfers' skill. *Journal of Sports Sciences, 21*, 593–600.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*, 29–46.

Phillips, L. (1988). The challenge of judgment. In J. Dowie (Ed.), *Professional Judgment and Decision Making, Introductory Texts 1*. Milton Keynes, England: The Open University.

Rassafiani, M., Ziviani, J., Rodger, S., & Dalgleish, L. (2008). Identification of occupational therapy clinical expertise: Decision- making characteristics. *Australian Occupational Therapy Journal, 55,* 1–11.

Seitz, K., & Schumann-Hengsteler, R. (2000). Mental multiplication and working memory. *European Journal of Cognitive Psychology*, *12*, 552–570.

Shanteau, J., Friel, B. M., Thomas, R. P., & Raacke, J. (2005). Development of expertise in a dynamic decision-making environment. In T. Betsch & S. Haberstroh (Eds.), *The routines of decision making* (pp. 251–270). Mahwah, NJ: Erlbaum.

Stewart, T. R., & Lusk, C. M. (1994). Seven components of judgmental forecasting skill: Implications for research and improvement of forecasts. *Journal of Forecasting, 13*, 579–599.

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Doubleday.

Uebersax, J. S., & Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine*, *9*, 559–572.

Weiss, D. J. (2006). *Analysis of variance and functional measurement: A practical guide*. NY: Oxford University Press.

Weiss, D. J., & Edwards, W. (2005). A mean for all seasons. *Behavior Research Methods, 37*, 677–683.

Weiss, D. J., Edwards, W., & Shanteau, J. (2009). The measurement of behavior: Indices and standards. In J. W. Weiss & D. J. Weiss (Eds.), *A science of decision making: The legacy of Ward Edwards* (pp. 262–268). New York: Oxford University Press.

Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors, 45,* 104–116.

Weiss, D. J., & Shanteau, J. (2004). The vice of consensus and the virtue of consistency. In K. Smith, J. Shanteau, & P. Johnson (Eds.), *Psychological investigations of competent decision making* (pp. 226–240). Cambridge, UK: Cambridge University Press.

Weiss, D. J., Shanteau, J., & Harries, P. (2006). People who judge people. *Journal of Behavioral Decision Making, 19*, 441–454.

Williams, C. A., Haslam, R. A., & Weiss, D. J. (2008). The Cochran-Weiss-Shanteau performance index as an indicator of Upper Limb Disorder risk assessment expertise. *Ergonomics, 51,* 1219–1237.