Technical Report 1153

Applying Consensus Based Measurement to the Assessment of Emerging Domains

**Peter J. Legree, Joseph Psotka,
and Trueman R. Tremble, Jr.**
U. S. Army Research Institute

Dennis Bourne
Howard University
Consortium Research Fellows Program

January 2005

20050322 120

United States Army Research Institute
for the Behavioral and Social Sciences

**U.S. Army Research Institute
for the Behavioral and Social Sciences**

**A Directorate of the Department of the Army
Deputy Chief of Staff, G1**

**ZITA M. SIMUTIS
Director**

Technical Review by

Paul A. Gade, U.S. Army Research Institute
Jennifer Solberg, U.S. Army Research Institute

NOTICES

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy) January 2005	2. REPORT TYPE Final	3. DATES COVERED (from. . . to) August 2003 – August 2004	
4. TITLE AND SUBTITLE Applying Consensus Based Measurement to the Assessment of Emerging Domains		5a. CONTRACT OR GRANT NUMBER	
		5b. PROGRAM ELEMENT NUMBER 0611101A	
6. AUTHOR(S) Peter J. Legree, Joseph Psotka, Trueman R. Tremble, Jr. (U.S. Army Research Institute), and Dennis Bourne (Howard University)		5c. PROJECT NUMBER 2O1611101A91E	
		5d. TASK NUMBER 296	
		5e. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 2511 Jefferson Davis Highway ATTN: DAPE-ARI-RS		8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 2511 Jefferson Davis Highway ATTN: DAPE-ARI-RS Arlington, VA 22202-3926		10. MONITOR ACRONYM ARI	
		11. MONITOR REPORT NUMBER Technical Report 1153	

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Subject Matter POC: Peter Legree (703/602-7947)

14. ABSTRACT *(Maximum 200 words)*: Situational judgment tests have been developed in the fields of Industrial/Organizational and Cognitive Psychology to predict performance and to evaluate theories of cognition. Production of these scales has usually required the opinions of subject matter experts to produce scoring keys or criterion data to compute empirically based standards. A simpler, elegant procedure is considered that allows examinee responses to be scored as deviations from the consensus defined by the response distributions of the examinee sample. This approach is termed "Consensus Based Measurement" and has been applied to validate scales in domains, such as Emotional Intelligence, that lack certified experts and well-specified, objective knowledge. Data are summarized demonstrating substantial convergence between situational judgment test scores computed using expert and examinee based scoring standards for which substantial expert and examinee data are available. The convergence indicates that examinee response distributions may be used to score situational judgment tests when expert responses are not available. Validity data for situational judgment scales that are scored with this approach are summarized.

15. SUBJECT TERMS
Consensus Based Measurement, Consensual Measurement, situational judgment tests, Emotional Intelligence

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified	Unlimited	33	Ellen Kinzer Technical Publication Specialist 703/602-8047

Applying Consensus Based Measurement to the Assessment of Emerging Domains

**Peter J. Legree, Joseph Psotka,
and Trueman R. Tremble, Jr.**
U. S. Army Research Institute

Dennis Bourne
Howard University
Consortium Research Fellows Program

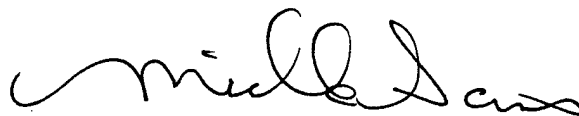**Selection and Assignment Research Unit
Michael G. Rumsey, Chief**

FOREWORD

This report describes "Consensus Based Measurement", the process of scoring items using a scoring standard derived from the responses of a large population of non-experts. This is a departure from traditional scoring methods that have utilized responses from a small group of experts or established facts to develop standards. The report uses as its context the assessment of Emotional Intelligence, a domain of research that is still in its formative stages. As such, Emotional Intelligence is a perfect vehicle for a discussion of consensus based measurement, because there are few, if any, experts and only a small base of concrete knowledge from which to develop scoring standards for test items.

The authors develop a variety of arguments for the use of consensus based measurement. First, traditional test construction and scoring methods are described in the context of their limitations to establish the requirement for an alternative scoring method. Then, the authors discuss theories of knowledge and expertise to provide an underpinning for the use of consensually derived standards using opinion data collected from non-experts. Next, a presentation of hypothetical and empirical data is provided to illustrate consensus based measurement, both in theory and in practice. These data establish the validity of scores derived using this method. Finally, there is a further discussion of the underlying theory and its limitations and implications.

With new arenas for scientific research and exploration emerging so often and so many existing arenas comprised of debated, and debatable, data and theory, this report offers an alternative to traditional methods of measurement. Consensus based measurement promises to increase the breadth of knowledge that may be directly measured to include domains lacking both experts and well-specified facts. Many of these domains may have direct relevance to Army operations and capabilities.

Ongoing projects at the United States Army Research Institute for the Behavioral and Social Sciences (ARI) are linking individual characteristics and aptitudes to outcomes that quantify job performance and Soldier attrition. Currently, measures being developed to assess knowledge specific to particular military occupational specialties (MOS) will be scored using consensus based measurement principles, and thereby, this approach may support Army personnel selection and classification policy. Moreover, this method can be used to develop measures to assess and develop the capabilities of recruits to provide effective peer support that mitigates factors associated with attrition.

MICHELLE SAMS
Technical Director

v

ACKNOWLEDGEMENTS

APPLYING CONSENSUS BASED MEASUREMENT TO THE ASSESSMENT OF
EMERGING DOMAINS

EXECUTIVE SUMMARY

Research Requirements:

Over the past decade, scenario based scales have been developed to measure knowledge, abilities, and expertise in performance domains. Most applications have utilized expert groups to develop scoring standards. Scales, and especially predictor scales, are produced using a subset of items selected to differentiate high and low performing examinees. Resultant tests are usually accurate, reliable, and frequently valid, against some external criterion.

However, much knowledge is intuitive and tacit, and might be called mere opinion, so there may be no formal knowledge sources, external criterion, or even experts who can provide appropriate standards. In many areas such as art, music, politics, government, and economics experts may have, or seem to have, markedly different views, rationales, and evidentiary sources than the diverse populations of interest to researchers.

Some applications of scenario-based scales have used scoring keys based on data collected from large groups of knowledgeable, but non-expert respondents. In these earlier papers, the use of non-expert groups to develop scoring standards was termed "Consensual Scoring" or more broadly, "Consensus Based Measurement (CBM)". CBM offers unique, analytic powers for exploration and measurement within domains that lack objective standards, an established body of knowledge or an available pool of experts.

Procedure:

The report compares CBM to traditional scale construction and scoring practices, details the methods and findings of a number of past efforts that have employed CBM as a scoring method, and provides a rationale for the use of CBM.

Findings:

The data summarized in this report demonstrate substantial convergence between situational judgment test scores computed using expert based scoring standard and those computed using examinee based standards: score correlations ranged from .88 to .995 across four applications for which expert and consensus based scores were available. This convergence indicates that examinee response distributions may be used to score situational judgment tests when expert responses are not available.

Comparisons of measures scored using examinee based standards with criterion measures showed a strong correlation between their respective scores. Findings such as this demonstrate that consensus based measures are both feasible and valid.

Utilization of Findings:

The assessment of knowledge corresponding to "soft" domains, or emerging domains such as emotional and social intelligence, where the codification and formalization of knowledge is only beginning, cries out for the use of this new technology. Though ill-defined, these domains are often of considerable consequence: knowledge and expertise related to driving safety, leadership, and social functioning can and do substantially impact an individual's performance and quality of life.

APPLYING CONSENSUS BASED MEASUREMENT TO THE ASSESSMENT OF
EMERGING DOMAINS

CONTENTS

List of Tables

List of Figures

INTRODUCTION

Over the past decade, scenario based scales have been developed to measure knowledge and expertise in performance domains such as leadership and driver safety, as well as to assess emotional and social intelligence, and general cognitive aptitude. While most applications have utilized expert groups to develop scoring standards (see Hedlund et al., 2003), other attempts have constructed scoring keys based on data collected from large groups of respondents who were knowledgeable concerning the subject domain but could not be qualified as experts. The scoring keys from these groups of non-experts were believed to have closely approximated the scoring standards that would have been obtained from experts. In these earlier papers, the use of non-expert groups to develop scoring standards was termed "Consensual Scoring" or more broadly, "Consensus Based Measurement (CBM)". CBM provides a maximal performance based method to assess knowledge-related constructs and is relevant to conceptualizations of emotional intelligence that propose a related set of knowledge, skills and abilities.

The promise of CBM rests in the fact that it expands the spectrum of knowledge addressed in psychological research to include domains for which neither bona fide experts can be identified, nor objective factual knowledge located. Consensus based measurement (CBM) is relevant to measuring emotional intelligence because it is an example of a domain that is still lacking in the availability of experts and objective knowledge. In fact, the theoretical development of emotional intelligence is still broadly viewed as in a stage of formative development. This fact notwithstanding, CBM has been used to score well-developed performance-based emotional intelligence scales, including the Multi-factor Emotional Intelligence Scale (MEIS; Mayer, Caruso, & Salovey, 1999) and the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT; Mayer, Salovey, Caruso, & Sitarenios, 2003). However, the notion that non-experts can be used to develop the "expert" knowledge required to score these instruments may be unappealing to test developers who are not yet familiar with the strengths and limitations of this approach, and those commentators who have questioned its assumptions (e.g., Roberts, Zeidner, & Matthews, 2001; Schaie, 2001; Zeidner, Matthews, & Roberts, 2001). Thus, a discussion describing CBM and its development in disparate areas of applied psychology (along with a summary of relevant data and theory) could help some favorable consensus to develop.

We will present a case for using CBM for ill-specified knowledge domains, such as emotional intelligence, and for other domains, where experts might not be available, because of some unique advantages associated with consensual scoring.

TEST CONSTRUCTION FOR POORLY SPECIFIED KNOWLEDGE DOMAINS

Many psychological knowledge tests are based on a job (or task or cognitive) analysis that associates knowledge and performance domains. Based on available data, this approach has proven its worth in many pragmatic areas of assessment and counseling (see Anastasi & Urbina, 1997). Implicit in this approach are expectations that formal and tacit knowledge underlie much performance, and that observed behavior supports inferences connecting behavior

with those knowledge attainments. Construction of knowledge scales traditionally has drawn either on an available, formal corpus of accumulated knowledge (such as books written by experts; or pedagogical materials developed over decades of instruction and analysis) or on an available pool of institutionally recognized experts.

However, much knowledge is intuitive and tacit, and might be called mere opinion, so there may be no formal knowledge sources, or even experts who can provide appropriate standards. In many areas, such as art, music, politics, government, and economics, experts may have, or seem to have, markedly different views, rationales, and evidentiary sources than the populations of interest to researchers. CBM offers unique, analytic powers in these situations.

Limitations of Traditional Scale Construction

While CBM may have some noteworthy limitations, we would like to point out that traditional item construction, based on de facto expertise, also has its limitations. Item construction in formal, well-defined knowledge domains can easily incorporate general knowledge and expertise, and item revision is often based on the use of item statistics or factor analytic techniques, to maximize scale characteristics such as reliability and validity. Because predictor and criterion reliability limit scale validity, the maximization of test reliability is of critical importance, and test construction decisions are frequently based on requirements to improve scale reliability. To maximize reliability, item statistics and especially low item correlations with total test score have been commonly used to identify questionable items for revision or deletion. From an Item Response Theory perspective, concerns with analogous goals result in the characterization of items as inefficient in providing information and requiring modification. Scales, and especially predictor scales, are produced using a subset of the items selected to differentiate high and low performing examinees. Resultant tests are usually accurate, reliable, and frequently valid against some external criterion.

For many academic and industrial purposes, this traditional approach has been adequate for the development of knowledge measures that are both valid and useful for personnel management and training decisions. Much mathematical knowledge, for example, is well developed and linked to performance, and it is relatively simple to identify the correct answer for a range of questions requiring the understanding of basic concepts. Likewise, words and expressions have specific meanings and connotations, as detailed in dictionaries. Vocabulary knowledge is frequently assessed with items corresponding to these dictionary definitions and is sometimes used to estimate general cognitive aptitude. Initial item construction is possible largely because of the presence of expert knowledge usually reflecting the availability of an information corpus and sometimes the opinions of experts. Even simple arithmetic and algebra problems require expertise, although it is widely available. The impact of the use of item statistics is to construct consistency within the measure, and create a stronger relationship between performance and the likelihood to respond correctly on all items. Seen from the perspective of CBM, this procedure in effect creates a consensus among the standardization group. From this perspective, all scales are consensually constructed, and consensus based scoring is a variant on a long established theme.

When Consensus Goes Awry

Obviously, items may occasionally be created for which consensus understandings are not correct or for which different groups have markedly different understandings: What is the capital of Israel (Tel Aviv/Jerusalem)? Should the US have invaded Iraq (Yes/No)? Where is the US federal government located (White House/Capital Building/Supreme Court/Executive Buildings)? These are all items for which different groups may have different understandings, or for which different understandings may have varying validity. A reasonable response to the presence of these occasional disagreements is not to reject CBM, but to understand the basis of these disagreements and thereby identify implications relating to the development and assessment of knowledge and opinion. Furthermore, the possibility that the knowledge underlying many questions might be deduced by analyzing the opinions of large numbers of non-experts is intriguing.

Knowledge Domains without Experts

It seems incontrovertible that knowledge domains may exist without the presence of an expert knowledge source, either in the form of an information corpus or verifiable experts. Consider that before the efforts of Noah Webster (1758-1843), assessing English language vocabulary knowledge of American colonists would have been problematic from the standpoint of scoring responses. Lacking a convenient information source for word knowledge (i.e., the dictionary), an 18th century vocabulary test developer might have felt compelled to determine, through expert opinions, acceptable definitions for American terms, such as "hickory", and for common terms that might have multiple meanings, such as "bed". Whether expert opinions would judge a flower garden reference as an acceptable definition for "bed" is an open question, but the direction of the judgment would impact individual scores.

But what population would constitute appropriate subject matter experts for the common English vocabulary knowledge of Webster's time? The use of highly regarded 18th century, United Kingdom English professors as subject matter experts might seem reasonable and would have foreshadowed approaches commonly used in Industrial/Organizational Psychology to develop and score situational judgment tests, but these opinions might have been skewed in an academic direction. It may be interesting that Noah Webster, who was also an important patriot dedicated to the democratic ideals of the American Revolution, incorporated definitions for uniquely North American terms such as "hickory" and "skunk". He also simplified spelling in a manner more consistent with Benjamin Franklin's preferences, substituting "center" for "centre" and "music" for "musick" (http://www.m-w.com/about/noah.htm). Royal English professors at Oxford and Cambridge Universities would seem unlikely candidates to accept these innovations, and this expected resistance would have produced questionable results if they were used as subject matter experts. It would be more reasonable (and Jeffersonian!) to survey a sample of English speaking American colonists/citizens and develop guidelines to identify acceptable responses for vocabulary definition items. In short, if we had to develop a vocabulary test today, without the benefit of dictionaries, using a democratic sampling of a broad spectrum of educated adults to act as experts would seem a reasonable approach.

This reasoning illustrates how knowledge domains may exist that are lodged in opinion and have no objective standard for verification other than societal views, opinions, and interpretations. Yet these knowledge domains may provide important information concerning one's abilities; after all, vocabulary knowledge has traditionally been very highly loaded on psychometric g (Carroll, 1993). For such knowledge domains, it may be mandatory to use standards based on a social knowledge perspective to evaluate individual responses. The concept that much knowledge is experientially based and linked to opinion is rooted in the writings of philosophers, such as Plato and John Stuart Mill. And, the concept that the opinions of common people may reflect higher standards is at the heart of democratic institutions.
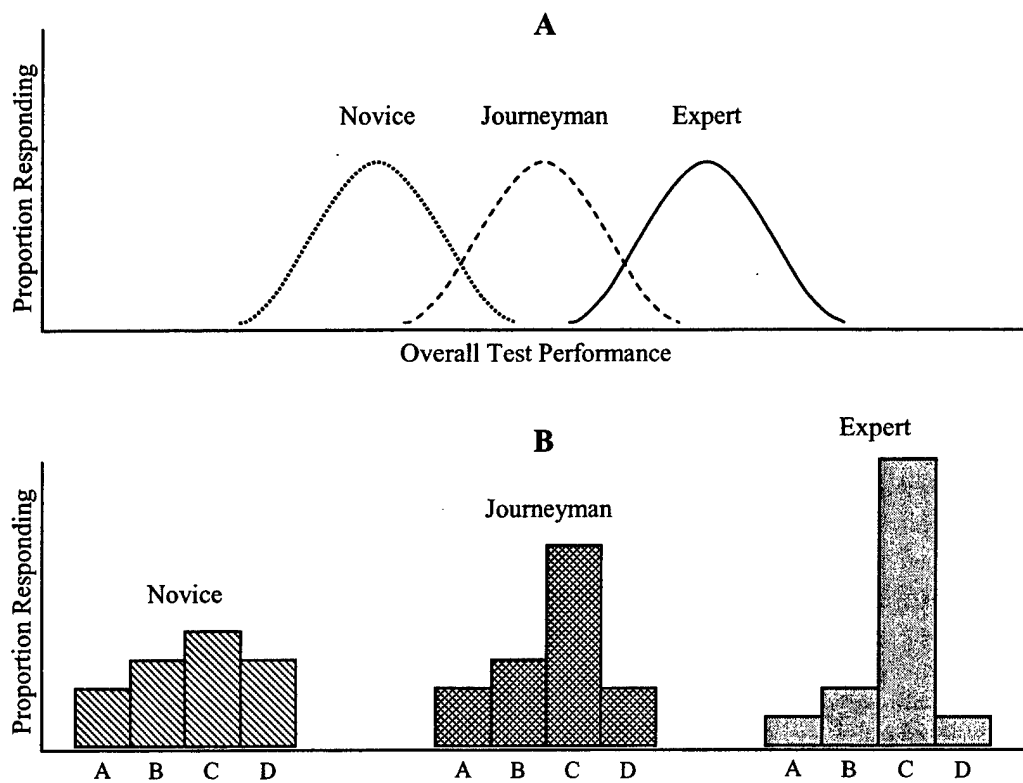
The assessment of knowledge corresponding to "soft", emerging domains such as emotional and social intelligence, where the codification and formalization of knowledge is only beginning, cries out for the use of these new technologies. These ill-defined domains are often of considerable consequence: knowledge and expertise related to driving safety, leadership, and social functioning can and does substantially impact on an individual's quality of life. It is important to this discussion that these knowledge domains are analogous to the situation that our 18th century vocabulary test developer would have experienced, because for these domains, well-developed knowledge corpora are not available and, equally important, identifying appropriate groups of subject matter experts is problematic. Scales developed to assess these domains might evaluate the consistency of an individual's cognitive structures with a scoring standard corresponding to a group consensus and therefore would be methodologically similar.

KNOWLEDGE, RESPONSE DISTRIBUTIONS, AND LEVELS OF EXPERTISE

Our conceptualizations regarding CBM evolved from expectations about how item response distributions might change as a function of the expertise of various respondent samples. Knowledge is customarily viewed as growing as levels of expertise increase in a specific domain. Therefore, if a sample of apprentices were tracked over time, and repeatedly surveyed with standard knowledge items as novices (or initiates), journeymen, and experts, the response distributions shown in Figure 1A might describe their growth in expertise. The distributions in Figure 1A illustrate both individual differences as well as increasing knowledge. For any individual test item, more respondents would choose the correct response as expertise increased, as is illustrated in Figure 1B.

However, suppose a sample of students studying EI were surveyed with items that required examinees to endorse statements on a Likert scale. For example, examinees might be requested to rate their agreement with the statement: "EI may be defined as the individual's fund of knowledge about the social world"; similar statements have been proposed to define Social Intelligence (see Cantor & Kihlstrom, 1987, but not Emotional Intelligence). For this type of question, the item response distributions associated with increased levels of expertise might vary in both central tendency and in variance. A change in central tendency might occur as students learn that some EI conceptualizations carry implications for social knowledge. Changes in the central tendency of these types of response distributions are illustrated in Figure 2A. A reduction in variance might also occur as students become more refined in their understandings of

A

Novice   Journeyman   Expert

Proportion Responding

Overall Test Performance

B

Expert

Journeyman

Proportion Responding

Novice

A B C D   A B C D   A B C D

Example item: "Which saw would you use to trim moldings?
(A) Azebiki saw, (B) Coping saw, (C) Dovetail saw, (D) Keyhole saw"

Figure 1. Performance on a conventional test at the scale and item level across three levels of expertise. Panel A: Overall scale performance distributions for a multiple-choice test. Panel B: Theoretical response distributions for a multiple-choice item where "C" is the correct answer.
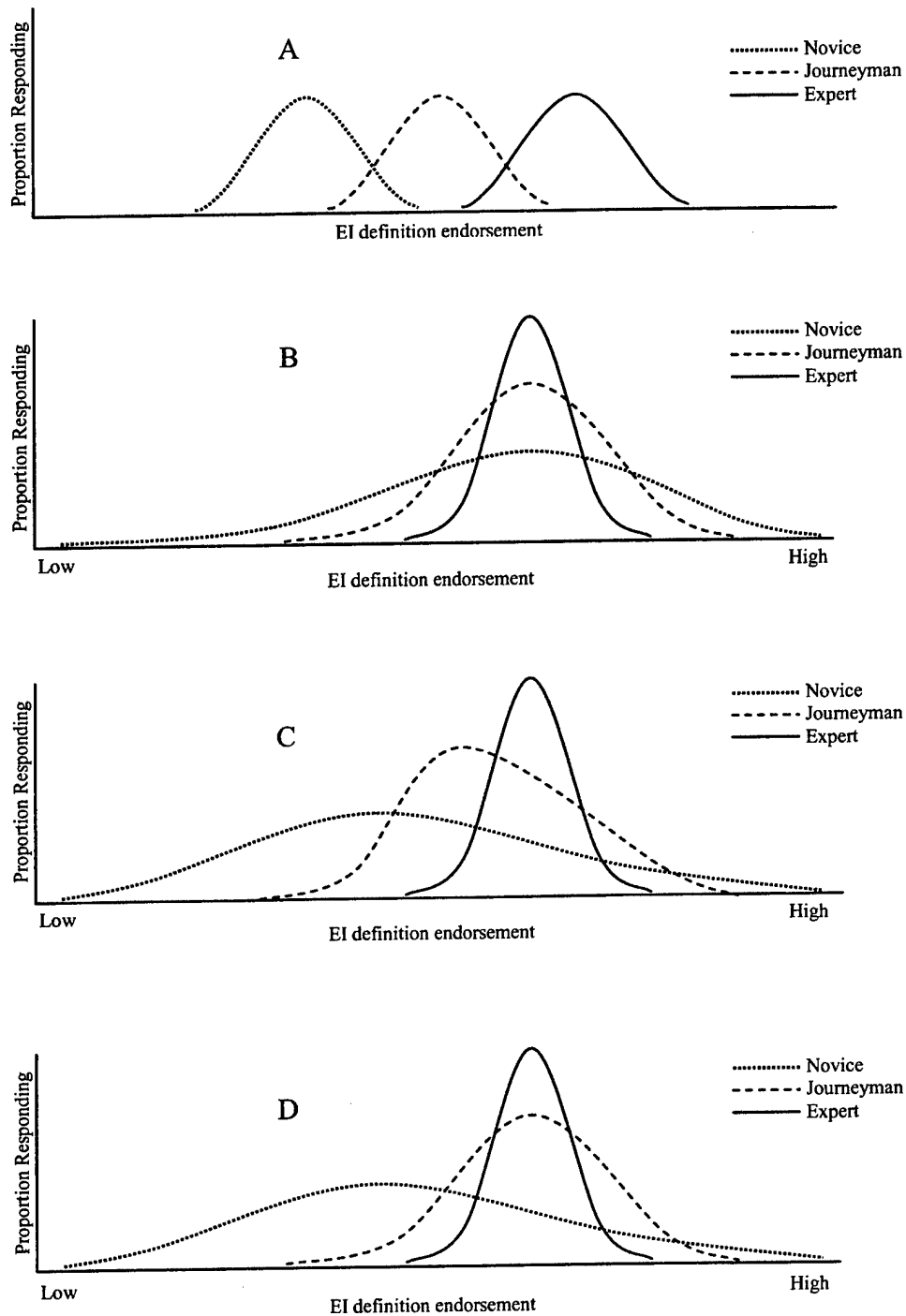
Figure 2. Hypothetical response distributions across three levels of expertise corresponding to a Likert-based item requiring endorsement of the statement: "EI may be defined as the individual's fund of knowledge about the social world". Panel A: Item distributions associated with changing central tendency and equal variance. Panel B: Item distributions associated with equal central tendency and differing variance. Panel C: Expected item distributions with differing central tendency and differing variance. Panel D: Observed item distributions for scenario-based items.

6

emotional intelligence, recognizing that while EI conceptualizations focus on emotion constructs, they also carry implications for social knowledge. Figure 2B illustrates a reduction in variance of response distributions associated with increased accuracy.

Both these trends have general relevance to understanding the growth and refinement of knowledge through reflection, experience, and formal education. By definition, naïve individuals have poorly formed conceptual structures for understanding relationships or events, and their responses may not be sensible, sometimes indicating ignorance of even basic relationships and sometimes overstating their importance. But with increasing degrees of sophistication, individuals will become increasingly aware and accurate in their understanding of relationships and events. It is worth considering that, to the extent poor performance on a knowledge test can be viewed as reflecting error, non-expert responses will be more variable than those of experts, as well as possibly having a different central tendency.

This conceptualization suggests that as error is reduced examinees will tend to agree with each other to a greater extent as expertise increases for both conventional and scenario-based test items. The central tendency of expert response distributions for individual, scenario-based items should be roughly equal to the central tendency of non-expert (e.g., journeymen) response distributions when the growth of knowledge over expertise is associated primarily with changes in variance (Figure 2B). This observation also applies to conventional multiple-choice items (Figure 1B), but it is of little practical value because writing sensible, multiple-choice items requires that the correct response be known a priori. Scenario based items do not always require that the correct response be specified or even known.

However, it is conceptually possible for increasing expertise to show changes in central tendency as well as variance. This model is intermediate and is represented in Figure 2C. At this time we have little meaningful to say about what kinds of items should show changes over levels of expertise in variance, central tendency, or both, with changing expertise, but simply point out the logical possibility and consider that a research agenda on CBM should investigate these relationships.

One powerful implication of successful CBM scales and inventories is a vindication and affirmation of broadly democratic processes that overturn the tyranny of autocratic expertise. Examining the hypothetical distributions of many novices (or initiates) against those of a handful of experts should reveal broader, flatter distributions that can more easily adapt and change with changing world knowledge. If one assumes that the correlation between novices' and experts' knowledge in these instruments is mediated by the intersection of their correlations with some broader truth, it may well turn out that diverse groups of novices may have a more accurate reflection of truth than experts; at least, this is a worthy hypothesis to investigate for limiting conditions. Some implications of these relationships are drawn below.

SITUATIONAL JUDGMENT TESTS AS MEASURES OF EMOTIONAL INTELLIGENCE

We recognized that situational judgment tests (SJT) are ideal for studying changes in item response distributions over expertise, in both central tendency and variance as described above. We describe situational judgment tests broadly as scales that:

1. Either implicitly or explicitly describe a scenario in order to simulate or depict an event, situation, or process. The scenarios may represent problems requiring solutions, the maintenance of success, or the interpretation of events. Understanding these depictions may require the application of knowledge gained either experientially or formally.

2. Provide a list of alternatives associated with each scenario. The alternatives may describe actions or interpretations, or provide the examinee the opportunity to respond in an open-ended manner to describe his/her opinion and knowledge.

3. Obligate examinees to either evaluate the alternatives associated with the scenarios (e.g., rating the appropriateness of the alternatives) or to generate new alternatives and analyses in the case of an open-ended response.

The scoring standards for most existent SJTs are developed by having subject matter experts evaluate or rate the alternatives for each scenario (see McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). These data are then used to construct the expert scoring standards, for example by computing mean expert ratings for each alternative. To evaluate examinee responses in comparison to the expert based standards, a percent-correct agreement, a deviation measure, or a correlation of an examinee's set of ratings with the scoring standard is computed. Consistent with the use of a variety of procedures to evaluate performance on SJTs with expert based scoring standards, various procedures might be used to consensually score SJTs . Information describing these possible methods is contained in Table 1. While these approaches have implicitly adopted a Classical Test Theory perspective, it is sensible that Item Response Theory analyses might be undertaken given sufficient data.
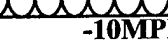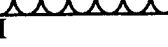
An SJT developed to evaluate tacit driving knowledge is presented in Table 2 to illustrate this approach. The driving knowledge test leveraged a model of driving performance recognizing that drivers may moderate risk by altering their speed in response to the presence of road hazards (Legree, Heffner, Psotka, Martin, & Medsker, 2003). This SJT was scored by computing distance scores between examinee responses and the scoring standard for each of the 14 items.

8

Table 1.

Approaches Used to Consensually Score SJTs

Method	Application	Citation
Percent Agreement: Likert data are collected and the most frequently obtained responses define "correctness." Then the approach is analogous to scoring standard dichotomous items.	Emotional Intelligence	Mayer et al. (2003)
Simple Distance: Likert data are used to compute item means over examinees. Distances are computed as the absolute difference between the individual and the mean rating for each item. Examinee performance is quantified as mean item distance.	Driving Knowledge	Legree, Heffner, Psotka, Martin, and Medsker (2003)
Standardized Distance: Similar to the Simple Distance method, but ratings are first transformed to standardize within individual. The approach controls for the tendency of some respondents to use only a sub-segment of the scale.	Social Intelligence & Psychometric g	Legree (1995); Legree, Martin, and Psotka (2000)
Squared Difference: Similar to Simple Distance, but item values are computed as the square of the difference. Provides additional weight to larger differences.	Tacit Knowledge	Sternberg et al. (2000)
Correlation: The value of the correlation of an individual's ratings with the mean ratings quantifies performance.	Leadership	Psotka, Streeter, Landauer, Lochbaum, and Robinson (2004)

Table 2.
Safe Speed Knowledge Test

Assume someone is driving a safe car in light traffic under optimal/perfect conditions. Given the following considerations, please <u>estimate</u> how much that individual (driver) should or shouldn't slow down and change speed to ensure safety.

CONDITIONS:	-20 MPH Slow Down	-10MPH	0 MPH Same Speed
1. Snow and heavy traffic			
2. Clear weather and light traffic			
3. Snow and no traffic			
4. Dry roads at midnight			
5. Stressed driver due to problems at work			
6. Moderately heavy traffic			
7. Gravel and light traffic			
8. Clear roads and somewhat breezy			
9. Light rain and curvy roads			
10. Angry and light rain			
11. Light traffic and hilly terrain			
12. Slightly worn tires			
13. Upset with family over finances/money			
14. Sick with a head cold			
	-20 MPH Slow Down	-10MPH	0 MPH Same Speed

Most SJT's have been produced for application within organizations. These scales usually present job-related problem scenarios and instruct examinees to choose among possible solution actions. In contrast, scales on the MEIS and the MSCEIT (Mayer et al., 1999; Mayer et al., 2003), and arguably Conditional Reasoning Tests (see James, 1998) present information and instruct examinees to choose among possible interpretations. Thus, the scales on the MEIS and the MSCEIT may be considered more abstract than standard SJT measures because they explore cognitive processes underlying behavior, as opposed to simply simulating observable decisions. Because these scales consider the basis of decisions, as opposed to being focused on the immediate results of decisions, we consider these scales to be assessing more abstract understandings and conceptualizations.

It is relevant to the current discussion that SJTs might be developed in other nontraditional manners to elucidate additional aspects of EI (or any other construct). For example, an SJT might be constructed that presents information and then estimates the time required for examinees to evaluate simple statements, or non-verbal stimuli, in a manner analogous to reaction or inspection time tasks (see Detterman, Caruso, Mayer, Legree, &

Conners, 1992). Such an SJT would measure processing speed associated with EI cognitions and would be consistent with emerging conceptualizations regarding chronometry (see Jensen, 1998).

Table 3.
Safe Speed Knowledge Item Response Distributions and Factor Loadings

Test Items	*M*	*SD*	% Speed[a]	Factor Loadings		
				Emotional Knowledge	Uncomplicated	Precipitation
Upset w/family finances	8.39	5.30	16	**.73**	-.01	.04
Sick with a head cold	8.50	5.08	13	**.73**	-.09	-.05
Slightly worn tires	7.62	4.92	3	**.55**	.05	-.02
Stressed over work	7.61	4.90	18	**.46**	.06	.17
Light traffic & hilly	6.17	4.28	20	**.44**	.17	.05
Clear & light traffic	1.50	3.19	78	-.02	**.92**	-.11
Clear & breezy	2.77	3.52	49	-.02	**.68**	.14
Dry & midnight	4.52	3.79	28	.20	**.62**	.02
Light rain & curves	10.40	4.07	1	.06	-.08	**.59**
Angry & light rain	10.59	4.43	3	.23	-.12	**.44**
Snow & no traffic	11.17	4.11	2	.02	.01	**.49**
Snow & heavy traffic	14.81	4.01	0	-.07	.13	**.40**
Mod. Heavy traffic[b]	7.70	4.17	8			
Gravel and light traffic[b]	7.77	4.02	7			

	Criteria Correlations			Factor Correlations		
	Fault Rate	Fault Status[c]	g			
Emotional	-.19✦	-.20✦	.10*	1.00	.19	.50
Dry Weather	-.10*	-.16†	.31✦		1.00	.25
Precipitation	-.16✦	-.16†	.11†			1.00

Note. n = 387. [a]Percent of respondents reporting that speed should not be reduced. [b]Variables excluded from analysis because of cross-loadings. [c]Fault status reflected an n of 211.
* p < .1. † p < .05. ✦ p < .01.

CBM: EMPIRICAL FINDINGS

Supervisory and Social Intelligence SJT Data

In earlier work with SJTs (Legree, 1995; Legree & Grafton, 1995), we evaluated our conceptualization of knowledge development by comparing expert based scoring standards that reflected the opinions of a small number of subject matter experts (i.e., mean expert ratings) and the mean ratings for the items as computed across examinees. That Supervisor SJT described 49 scenarios and listed a total of 198 alternatives, with between 3 and 5 alternatives per scenario. Each scenario described an interpersonal problem and presented alternatives as possible solutions to the problem. The scale was administered to examinees and experts who rated the appropriateness of the actions described in the alternatives for each scenario. We computed mean examinee ratings for each of the 198 alternatives and observed a high correlation between

11

the expert based scoring standard and the mean examinee item ratings, .72, (N=198, p < .001), and estimated a very high correlation, .95, by correcting the observed correlation for attenuation of the reliability of each set of observations (i.e., the mean expert and examinee ratings).

Initially, we had expected that examinee means would provide only a rough approximation of the expert-scoring standard. We had hoped this approximation would be evidenced by a moderate correlation between the means that would range between .4 and .6. We had planned to use a recursive procedure to sequentially identify groups of individuals with increasing levels of knowledge and then apply this approach to score scales for which expert opinions were not available. Information from the more select groups of individuals would then be used to develop increasingly valid scoring standards for the Supervisory SJT that would more closely approximate the expert standards. These standards would then be referred to as "consensus based standards" and the process as CBM.

Based on the observed and corrected correlations between the examinee and expert means, .72 and .95, the use of recursive procedures to refine the scoring pattern defined by the entire group of examinees was judged as not necessary. We also computed examinee scores using two different standards based on the expert and examinee means and then correlated the two sets of scores; this correlation was .88, (N=198, p < .001).

These correlations indicated that the mean ratings of examinees might provide an alternate scoring standard for the SJT, and this realization raised issues concerning the appropriateness of the two standards. We concluded that the examinee-based standard was preferable because these values were more reliable than the expert based standard, due to the large number of individuals (N=193). We then applied this method to score two additional social intelligence scales for which expert opinions were not available. A confirmatory factor analysis of these three scales and a standard aptitude battery (the Armed Service Vocational Aptitude Battery), demonstrated the existence of a separate g-loaded factor corresponding to our social intelligence model (Legree, 1995).

Applications to Assess g and Driver Safety

While the social intelligence model was confirmed, we recognized that the value of CBM needed to be buttressed in other domains by validating consensus based scores against conceptually relevant and important criteria, and by showing correspondence between scores based on examinee and expert opinions. In further research, we explored the power of CBM by developing and validating two types of scales: six Unobtrusive Knowledge Tests (UKTs), constructed to measure general cognitive aptitude, i.e., psychometric g, and two Tacit Driving Knowledge Tests, developed in order to assess knowledge related to driver safety.

With one exception, these measures required individuals to respond to items using Likert scales; for example, estimating the frequency of words and terms used in oral communication or the extent to which drivers should moderate speed when confronted with driving hazards. Construction of these scales leveraged conceptualizations of incidental learning and tacit knowledge to predict and understand human performance. This type of knowledge and associated expertise is usually acquired slowly and incrementally as a result of experience and

reflection upon those experiences (Sternberg et al., 2000). For these scales, neither an objective knowledge base nor experts could be identified to develop scoring standards. Thus, performance on these scales could only be evaluated using consensus based scoring algorithms.

The UKT battery was administered to a highly selected military sample comprised of Air Force recruits. Factor scores extracted from the UKT battery correlated .54 with factor scores extracted separately from the Armed Service Aptitude Battery (ASVAB); and a .80 correlation estimate was obtained by correcting for range restriction (Legree et al., 2000). This parameter estimate of .80 is typical of correlations obtained among cognitive test batteries (see Carroll, 1993). A LISREL confirmatory factor analysis of the corrected correlation matrix estimated a .97 path coefficient between the two latent factors corresponding to the Unobtrusive Knowledge and Conventional Test Batteries. Thus, we intentionally produced and scored a test battery highly correlated to a conventional test battery without using subject matter experts or objective knowledge, instead using CBM.

The tacit driving knowledge tests were administered to Army Soldiers, for whom automobile crash involvement data were also collected. Compared to most performance domains, crash involvement is unusual because it has only very minor relationships with knowledge, skill, and ability measures, including general cognitive aptitude, based on meta-analyses (Arthur, Barrett, &Alexander, 1991; Veling, 1982). However, as reported in Table 3, both of the tacit driving knowledge tests correlated significantly with crash involvement criteria, −.11 to −.20 (Legree et al., 2003). While these values may appear modest, they exceed coefficients typically obtained for stable characteristics and they carry implications for improving driver safety. Thus, the values we obtained demonstrate the utility of using consensus based scoring to assess tacit knowledge for this arguably atypical performance domain.

The Safe Speed Knowledge test, presented in Table 2, was one of two scales developed to assess tacit driving knowledge. Confirming the importance of constructs related to emotional intelligence, when the Safe Speed items were factored, one of the three factors was defined by emotionally and internally relevant items (see Table 3). Although this factor had a very minimal g loading, it was most predictive of the at-fault crash criteria. These data show safer drivers are more aware of the importance of moderating speed when under emotional (or internal) stress.

Of course individuals could have been nominated as "experts" to develop the scoring standards associated with the domains referenced by the UKT and tacit driving knowledge tests, but, it is our belief, all expert accreditations or knowledge corpora linked to these domains are suspect for their intended purposes. Nominated experts, having no more real expertise than the examinees in this study, would differ qualitatively, and not quantitatively, from the examinees who completed the scales. Knowledge of word frequency during oral communication and safe driving speed are exemplars of domains associated with experience that lack bona fide experts.

The implicit association task, which is the only experimental test that did not use a Likert response scale, is unique. The implicit association task assessed an examinee's ability to understand binary patterns (see Psotka, 1977), and each item required examinees to continue a series of X's and O's (e.g., XOXOXO?). No scoring standard could be invoked because the patterns used as stimuli were not chosen in accordance with pre-specified rules or relationships

13

that would dictate the correct answer. As a result, these items could only be consensually scored. Nevertheless, performance on this task correlated with psychometric g, $\rho = .40$ (Legree et al., 2000).

Additional Datasets Supporting Expert and Examinee Comparisons

The above data demonstrate the efficacy of CBM, for producing predictive validity, and by implication, for developing useful scoring standards. There is little doubt that CBM can be used to score tests developed for these unusual soft knowledge domains that lack formal sources of knowledge, which have either very high g, or minimal g loadings. Our conceptualization also predicted a very high correlation between expert and consensus based scoring standards, as well as the scores based on those two standards. For example, in our initial evaluation of the model, the expert and consensus scoring standards developed for the Supervisor SJT correlated .72, and scores based on these standards correlated .88. Because experts are often hard to find and expensive once found, much research with expert based measures has utilized small samples of experts and resulting scales and standards have had marginal levels of reliability. We are aware of three other data sets, discussed below, that used expert based standards derived from a large number of experts and are thus likely to have the needed level of reliability. There are likely to be additional datasets that could support these types of analyses, but examinee data are rarely used to approximate expert judgments. The available data are summarized to indicate the generality of the results.

Noncommissioned Officer (NCO) SJT. The largest of these data sets corresponds to the non-commissioned Officer (NCO) SJT developed to evaluate supervisory skills for senior enlisted Soldiers. The NCO SJT described 71 problem scenarios and listed 362 actions. To evaluate CBM, response protocols were scored using both expert (N=88) and consensus (N=1891) based standards (Heffner & Porr, 2000, W. B. Porr, personal communication, July 2003). Overall performance scores correlated .95 and scoring standards correlated .89.

Emotional Intelligence Data. The MSCEIT (Mayer et al., 2003), which is arguably the best-developed performance emotional intelligence battery, provides both expert and consensus based scores. The expert group consisted of 21 members of the International Society for Research on Emotions, and the consensus scores were derived from the responses of 2112 examinees, all of whom completed the scale. The correlation between the scores based on the two sets of standards was .98 and the score standards correlated .91. These researchers also reported inter-rater kappa coefficients for the experts and for two samples of the non-expert examinees: expert kappas were consistently higher than the examinee kappas (.43 versus .31/.38, $p < .01 / p < .05$) as suggested by a model of decreasing variance with increasing expertise, while central tendency remains constant.

Tacit Knowledge for Military Leadership Data. The third database corresponds to the Tacit Knowledge for Military Leadership (TKML) scale (Hedlund et al., 2003; Psotka et al., 2004). The TKML was designed to measure the practical, action-oriented knowledge that Army leaders typically acquire from experience. The TKML was developed with the idea that an ordered hierarchy of expertise in Military Leadership can be created by using the scores of Lieutenant Colonels as a standard and comparing them with U.S. Military Academy (West Point)

Cadets, and U.S. Army Lieutenants, Captains, and Majors. The scale was administered to groups of Soldiers including: 355 cadets, 125 lieutenants, 117 captains, 98 majors and 50 lieutenant colonels. The lieutenant colonels comprised the expert group, and this group contains the highest ranking Soldiers and those who have served longest in the military (with an average of 18 years of service). Comparisons of the consensus based cadet (355 cadets) and expert (50 lieutenant colonels) scoring standards and scores provide very consistent results with the earlier data. The two sets of score standards correlated .96, and the two sets of cadet scores computed using those standards correlated .995. Similar results were found by analyzing the data for the intermediate (lieutenant, captain, and major) groups.

While obtaining high correlations between the expert based and consensus based standards helps validate the approach, values approaching 1.00 were unexpected. In addition, the use of recursive procedures to refine consensus based standards was not required for the scales we developed. Collectively, these findings suggested that modification to our conceptualization of the CBM model might be warranted such that the principal difference between journeymen and experts is represented in terms of increasing accuracy, or from the perspective of item response distributions, decreased variance around the item means. However, the transition from novice to journeyman might still be associated with shifts in response distributions and means especially when novices have little or no basis for their responses and their responses are highly random. This revised model is represented in Figure 2D. To evaluate this model, it is necessary to inspect the response distributions of sizeable samples of individuals from groups varying in level of expertise.

Most databases are not adequate for this purpose because in most non-stratified samples there are very few novice or expert performing individuals, and identifying these individuals would be difficult. However, the TKML database is unique because it contains substantial numbers of novices (355 cadets), experts (50 lieutenant colonels) and examinees at the journeyman levels (125 lieutenants, 117 captains and 98 majors). These groups differ on a number of salient dimensions that affect expertise: age, experience, and education. In fact, cadets have really very little experience of the Army, but they do have some experience with interpersonal events and problems, and issues of authority, caring, and obedience that underlie the scenarios in the TKML; so although they are novices, they do have pertinent knowledge.

It should not be too surprising, then, that when the means of 355 cadet item response distributions were correlated with the means of 50 lieutenant colonels (experts), the overall correlation was quite high (r =.96) and the slope was close to one (0.99). The slope indicates a similar level of variance across the two sets of item means. Thus, despite the difference in expertise between cadets (with very little experience) and lieutenant colonels (with an average of 18 years of military experience), the use of the group's average as the standard is indistinguishable from an expert based score. And yet, the same standard still cleanly discriminates between these two groups. Although the overall item mean for each of the scenario alternatives was practically the same for cadets and colonels, even the top 25% of cadets scored significantly lower than the colonels on the overall TKML scale. Overall, the mean of the top 25% of the cadets was 0.73, whereas the colonels' mean score was 0.82(t = 4.27, 132 df, p < .01), which is equivalent to a difference of 0.36 standard deviation units, in favor of the colonels,

demonstrating that consensus based standards are effective in assessing what the scale was intended to assess: military leadership knowledge.
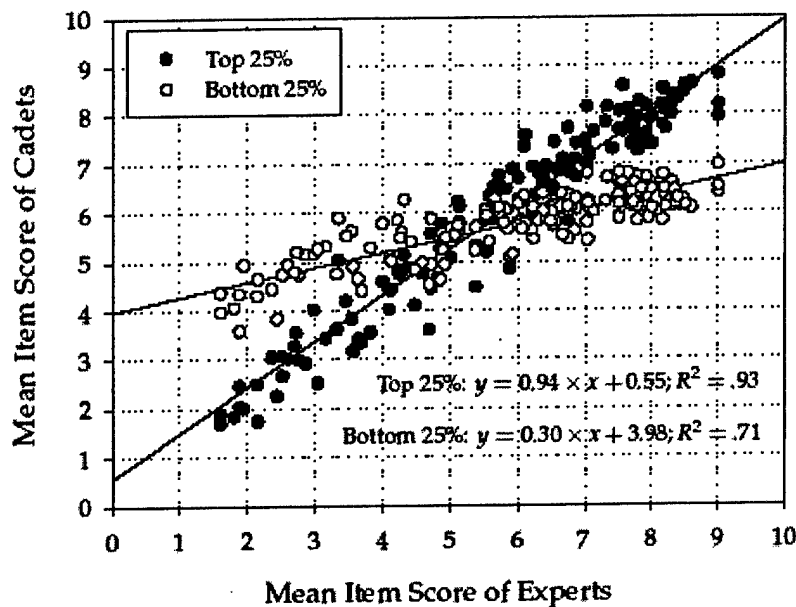
Figure 3. The relationship between the top 25% of Cadets, the bottom 25%, and the Expert Senior Officers used to standardize the TKML, showing that the top 25% are practically indistinguishable as a group for setting the standards of the test.

Differences between scoring standards can be demonstrated using the TKML dataset, but only by isolating a group associated with a very low level of expertise and comparing their means with values based on the other groups. Figure 3 shows exactly this sort of difference between the top and bottom 25% of the cadets at USMA.

For the top 25%, the correlation with the experts is .95 and the slope is 1.00. But for the bottom 25%, the correlation with the experts is .85 and the slope is 0.31. The low slope indicates less variance in item means computed using the lowest 25% of the cadets. Only by artificially restricting the examinee sample to the lowest quartile of the cadet sample can substantial changes in the standards be effected, and even then, the correlation is still .85.

If our notions of how expert knowledge is measured in these consensual scales are accurate, not only should novices have a lower correlation with experts, than journeymen with intermediate levels of expertise, but the slope of the regression line should also be lower. To understand this prediction, think of how the many different, and less correct, opinions of novices should combine. In the absence of systematic biases, components of the novices' thinking should be in error in different ways, but the components that are on the "road" to expertise should be similar. As there is more and more error, the overall regression to the mean should be stronger and stronger, giving rise to lower slopes. Thus, the TKML data are most consistent with a model in which experts and journeymen differ primarily in variance, with changes in central tendency

16

being more closely related to differences between novices and journeymen; this model is illustrated in Figure 2D.

CONCEPTUALIZING CBM: TOWARDS A WORKING MODEL

To describe CBM and summarize data describing its effectiveness and utility were two goals of this chapter. But the initial model was more descriptive than theoretical, and the concept that expert knowledge can be approximated by surveying large numbers of non-experts must have some limitations. So a more theoretical explanation of CBM is warranted. To understand consensus based scoring, it is useful to consider that for most knowledge domains, and especially for procedural knowledge domains, knowledge accumulates as the result of experience (see Anderson & Lebiere, 1998). As a greater range of events is experienced, greater levels of knowledge and associated skills will be acquired, and reactions to a new event or situation may reflect increasing levels of sophistication.

When presented with a situation to analyze, novices will have little basis for their opinions, and they will frequently disagree among themselves as well as with experts. Disagreement among novices is expected because the knowledge and cognitive structures associated with an individual novice will reflect either the action of a few unique experiences or the actions of experiences that have marginal relevance to the depicted situation. Thus, novices will reference different experiences and expectations, and their opinions will tend to be inconsistent, both among themselves as well as with experts.

In contrast, experts will generally have well-developed, mature knowledge structures reflecting broad, extensive sets of experiences. While each expert will have a slightly different set of experiences, these sets will largely overlap across individual experts. Moreover, with increasing levels of expertise, knowledge structures and related opinions will become progressively more consistent. Journeymen with partially developed and varying levels of expertise will agree at a moderate level both among themselves and with experts, and this moderate level of agreement is based on developing cognitive structures that reflect a modest but not extensive array of experience. From a mathematical perspective, the correlation of knowledge between individual A and individual B can be conceptualized as the product of the correlation of individual A with the 'truth' and of individual B with the 'truth'. As individuals A and B become more knowledgeable and their opinions more "truthful", their opinions and responses will become more highly correlated (see Romney & Weller, 1984).

In theory this progression is reasonable, but in some domains experts frequently disagree with each other, and expert performance may not be very impressive in comparison to non-expert performance: Clinical psychology, graduate admissions, and economic forecasting are all examples of domains in which it has been suggested that experts do not perform much better than novices (Chi, Glaser, & Farr, 1988). Thus, expectations concerning levels of expert agreement may easily be overstated and a more realistic perspective is to expect experts to differ quantitatively and not qualitatively from journeymen. In fairness to experts in these domains, these individuals may perform better, i.e., more consistently, than novices.

17

Because procedural knowledge is experientially based and because these experiences are dependent on the occurrence of real-world events, various journeymen may have different types of experiences and knowledge, although much of this knowledge will be most relevant to those situations that frequently occur. It follows that the breadth of experience associated with a single expert, while more extensive than that of an individual journeyman, will often be exceeded by the variety of experiences associated with a substantial number of journeymen. The implication of this view for CBM, as well as for other knowledge engineering applications, is that more information might be present in the knowledge structures of a large number of journeymen than a small number of experts.

The concept that expertise represents the sum total of many small components relates well to theories of intelligence to the extent that intelligence can be viewed as reflecting general life expertise. Thomson (1928, 1939) conceptualized psychometric g as arising from the separate action of many connections that sum to represent one's level of intelligence, and this view is based on the application of sampling theory to the measurement of intelligence. Under this model of intelligence, no single individual would perform perfectly across all connections, but across individuals, all connections would occasionally be closed. IQ tests were viewed as sampling these connections to estimate one's overall level of connectivity or general intelligence. A high IQ would evidence a high proportion of connections, and a low IQ would evidence a low proportion. However, low and moderate IQ scores could easily result from separate and sometimes non-overlapping sets of connections, for example when different individuals are knowledgeable regarding facts in different domains but unable to respond to many other queries. In modern parlance, g might be viewed as the sum of a very large number of separate factors or cognitive structures.

Because learning theories associate knowledge and experience, Thomson's view of intelligence as representing the sum of many small parts or connections has relevance. Expertise can be conceptualized as reflecting one's overall number and strength of cognitive structures; just as cognitive aptitude might reflect the presence of connections. Across individuals, lower levels of expertise can reflect cognitive structures that are largely non-overlapping sets of events, with higher levels of expertise reflecting more complete sets of cognitive structures and experiences. As in Thomson's analysis, no single individual can be expected to have experienced the universe of events associated with a domain of expertise. However, a large number of individuals, each with a moderate level of experience, could be expected to experience most, if not all, classes of events and to have cognitive structures correspondent with those events.

These learning theories are most relevant to understanding CBM when cognitive structures and related knowledge reflect the experience of largely unpredictable events, as does much procedural and tacit knowledge. In contrast, academic knowledge reflects more formal instruction, which is often structured to provide a systematic, highly ordered set of experiences based on objective information, and the surveying of students on topics not yet covered is unlikely to identify much information. However, all of the domains described in the current chapter correspond to incidental, tacit, or procedural knowledge. With respect to the SJT methodology (e.g., see McDaniel et al., 2001), a similar set of conditions prevail, as appears the case, we suspect, in many soft, poorly defined domains of psychological inquiry.

Thus, cognitive theories related to the acquisition of procedural knowledge support the contention that the opinions of a large number of journeymen can be used to approximate those opinions of a smaller number of experts for these types of domains, and this notion is the heart of CBM. In this chapter, we addressed the utility of CBM for scenario-based scales, on which examinees might respond on Likert scales. It is important that our results are consistent with simulations using dichotomous items when examinees are available but the objective answers are not specified (Batchelder & Romney, 1988). These analyses show highly accurate answer keys can be constructed using relatively small sets of respondents with the number of respondents in balance with the expertise of the group. These data also show that a majority rule may be used to infer correct responses given a large number of respondents. Of course it is rare for this procedure to be required for a dichotomous scale developed for a conventional knowledge domain, but these results are entirely consistent with our findings and the conclusion that CBM is ideal for poorly specified, emergent knowledge domains. It seems likely that this approach will remain relevant to developing scales for emerging domains, especially those based on experience, such as emotional intelligence, until these emerging domains become much better specified.

IMPLICATIONS FOR CONSENSUS BASED MEASUREMENT

Consensual scoring has several important implications for studying individual differences. First, the approach allows the construction and scoring of scales for knowledge domains for which experts do not exist, or cannot be easily identified. This allows an expansion of the domains for which knowledge tests may be developed, an expansion beyond traditional formal domains into everyday knowledge areas that are meaningful and important in our daily lives. Thus, consensus based scoring allows the assessment of knowledge domains that have not been traditionally addressed in psychological or educational research, and broadens the domain of psychological assessment and intelligence research into horizontal aspects of cognitive aptitude, one of which may be emotional or social intelligence. This perspective is consistent with theories of implicit and tacit knowledge acquisition and relates well to conceptualizations of social knowledge.

A second important implication is that CBM provides economy to test development. The approach allows questions to be posed, answered, and scored without the correct responses known a priori. Thus, the scale development cycle is shortened because expert responses are not required to construct scoring standards. In addition, costs associated with the production of scoring standards and rubrics are minimized because expert judgments can be expensive to collect while the examinee data are incidental to scale administration. A related implication results from the use of the Likert format to support CBM. Likert scales allow distances to be computed at the item level, thus allowing a more complete analysis of available information. As might be expected based on the use of additional information, comparisons of scores based on the distance information versus those based on a dichotomous format associates higher levels of reliability with the distance-based measures (Legree, 1995), and therefore the Likert format supports improved testing efficiency. In addition, distance items can be correlated, and factors extracted from these correlations have been sensible (see Legree et al., 2003).

Third, consensus based scoring has the potential to allow the same protocol to be scored against multiple standards. This approach could be useful in studying controversial domains associated with groups that may adopt different perspectives. This approach might relate well to understanding controversial views differing over gender, political affiliation, race, age, or sexual orientation or in identifying the basis for competing theories to explain some phenomenon. Who knows, it might even be applied to the development of formal theories of scaling and item measurement, such as consensual scaling, in an interesting recursive cycle!

Fourth, consensual scoring explicitly invokes the notion of disagreement and inconsistency in the coherence of knowledge structures. Ill-defined domains are characterized by disagreement even among experts. Factorial analysis and multidimensional scaling of their responses (Psotka et al., 2004) using such powerful technologies as Latent Semantic Analysis not only brings order to these disagreements, but provides the prospect of being able to define the source of differences and create new conjunctions in the informal frameworks. To paraphrase an oft-cited opinion[1]: "An intuitive inconsistency is the muse of great minds."

Fifth, consensus-based scoring emphasizes that under at least some conditions, standards based on a body of relatively informed individuals approximate the standards of experts. SJTs are sometimes called "low fidelity simulations" because they provide the minimum stimulus cues needed to evoke responses representing the phenomenon targeted for measurement. As such, SJTs present somewhat ambiguous situations. Our principal interpretation suggests that judgments to these ambiguous situations are direct reflections of existing knowledge. A complementary explanation inspired by Gestalt psychology, is that abstract stimulus situations do not create all cues needed for a response, instead forcing interpretation or induction of meaning. Thus, rather than directly reflecting the qualities of existing knowledge structures, responses reflect the existing structures mediated by the understanding reached about the abstract situations. Superior performance would then reflect greater access to commonality in forcing interpretation and induction of meaning.

Under conditions when paradigm shifts develop or when information is distributed that differentially influences either expert or journeyman opinions, or when these conditions result in group divisions that retard rather than further group goals, then it seems less likely that CBM will produce a useful metric of group agreement needed to evaluate expertise. Whether a multi-modal approach could be used to develop multiple metrics is an open question, but this approach might have relevance to understanding interactions between groups that sometimes conflict.

Much social knowledge represents the convergence between many perspectives and truth is commonly believed to exist at the intersection of these perspectives. Thus, the American legal system, with one side designated as prosecutor and the opposing side as defendant is a manifestation of this view, as are all democratic institutions. The perspective that knowledge is rooted in widely diverse opinion is reflected in Tolstoy's observation that "Happy families are all alike; every unhappy family is unhappy in its own way", and from a cross-cultural perspective, the African proverb, "It takes a village to raise a child". The success of these institutions and the relevance of these statements reflect the notion that knowledge can be distributed over individuals, and would appear consistent with use and development of technologies to identify

---

[1] "A foolish consistency is the hobgoblin of small minds" – Ralph Waldo Emerson.

this type of knowledge and its evidentiary sources for emerging fields such as social and emotional intelligence.

EPILOGUE

We would like feedback from our readers. Using a 9-point Likert scale, please email your ratings of the extent (1=not at all...9=completely) to which you believe:

1. You are very knowledgeable concerning test development.
2. Traditional test development methods are appropriate for well-specified knowledge domains.
3. Traditional test development methods are appropriate for emerging, ill-specified knowledge domains.
4. CBM methods are appropriate for well-specified knowledge domains.
5. CBM methods are appropriate for emerging, ill-specified knowledge domains.
6. Academic knowledge can be accurately measured using multiple-choice measures.
7. Academic knowledge can be accurately measured using Likert based items.
8. Procedural knowledge can be accurately measured using multiple-choice measures.
9. Procedural knowledge can be accurately measured using Likert based items.
10. It is reasonable to expect that happy families are more similar than unhappy families.

If we collect sufficient information, we will compare the response distributions of readers of this chapter for these items to those collected from test-developers who have not reviewed this information. If our theory is correct, then a greater level of agreement for CBM related items should be apparent for the chapter readers than for the non-readers as evidenced by decreased variance yet similar means over those items. Please respond to the first author by email, legree@ari.army.mil.

Author Note

The views, opinions, and/or findings contained in this article are solely those of the authors and should not be construed as an official Department of the Army or DOD position, policy, or decision, unless so designated by other documentation.

REFERENCES

Anastasi, A., &Urbina, S. (1997). *Psychological testing*. Upper Saddle River, NJ: Prentice-Hall.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum.

Arthur, J.,W., Barrett, G. V., & Alexander, R. A. (1991). Prediction of vehicular accident analysis: A meta–analysis. *Human Performance, 4*, 89–105.

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

Cantor, N., & Kihlstrom, J. F. (1987). *Personality and social intelligence*. Englewood Cliffs, NJ: Prentice-Hall.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor–analytic studies*. New York, NY: Cambridge University Press.

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum.

Detterman, D. K., Caruso, D. R., Mayer, J. D., Legree, P. J., & Conners, F. (1992). Assessment of basic cognitive abilities in relation to cognitive deficits; mopping up: Relation between cognitive processes and intelligence. *American Journal on Mental Retardation, 97*, 251–286.

Hedlund, J., Forsythe, G. B., Horvath, J. A., Williams, W. M., Snook, S., & Sternberg, R. J. (2003). Identifying and assessing tacit knowledge: Understanding the practical intelligence of military leaders. *Leadership Quarterly, 14*, 117–140.

Heffner, T. S., & Porr, W. B. (2000, August). Scoring situational judgment tests: A comparison of multiple standards using scenario response alternatives. Paper presented at the Annual Conference of the American Psychological Association, Washington, DC.

James, L. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods, 1*, 131–163.

Jensen, A. R. (1998). The g factor. Westport, CT: Praeger Publishers.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor. *Intelligence, 21*, 247–266.

Legree, P. J., & Grafton, F. C. (1995). *Evidence for an interpersonal knowledge factor: The reliability and factor structure of tests of interpersonal knowledge and general cognitive ability* (ARI Technical Report No. 1030). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Legree, P. J., Heffner, T. S., Psotka, J., Martin, D. E., & Medsker, G. J. (2003). Traffic crash involvement: Experiential driving knowledge and stressful contextual antecedents. *Journal of Applied Psychology, 88*, 15–26.

Legree, P. J., Martin, D. E., & Psotka, J. (2000). Measuring cognitive aptitude using unobtrusive knowledge tests: A new survey technology. *Intelligence, 28*, 291–308.

Mayer, J. D., Caruso, D. R., & Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence, 27*, 267–298.

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Modeling and measuring emotional intelligence with the MSCEIT V2.0. *Emotion, 3*, 97–105.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740.

Psotka, J. (1977). Syntely: Paradigm for an inductive psychology of memory, perception, and thinking. *Memory and Cognition, 3*, 553–600.

Psotka, J., Streeter, L. A., Landauer, T., Lochbaum, K. E., & Robinson, K. (2004). Augmenting electronic environments for leadership. In *Proceedings of the human factors and medicine panel, Genoa, Italy.*

Roberts, R. D., Zeidner, M., & Matthews, G. (2001). Does emotional intelligence meet traditional standards for an intelligence? some new data and conclusions. *Emotion, 1,* 196–231.

Romney, A. K., & Weller, S. C. (1984). Predicting informant accuracy from patterns of recall among informants. *Social Networks, 6*, 59–77.

Schaie, K. W. (2001). Emotional intelligence: Psychometric status and developmental characteristics – comment on Roberts, Zeidner and Matthews. *Emotion, 1*, 243– 248.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., Snook, S., & Grigorenko, E. L. (2000). *Practical intelligence in everyday life.* New York, NY: Cambridge University Press.

Thomson, G. H. (1928). A worked out example of the possible linkages of four correlated variables on the sampling theory. *The British Journal of Psychology, 18*, 68–76.

Thomson, G. H. (1939). *The factorial analysis of human ability.* New York, NY: Houghton-Mifflin Company.

Veling, I. H. (1982). Measuring driving knowledge. *Accident, Analysis, & Prevention, 14*, 81–85.

Zeidner, M., Matthews, G., & Roberts, R. D. (2001). Slow down you move too fast: Emotional intelligence remains an elusive intelligence. *Emotion, 1*, 265–275.