



ELSEVIER

European Journal of Operational Research 000 (2001) 000–000

EUROPEAN
JOURNAL
OF OPERATIONAL
RESEARCH

www.elsevier.com/locate/dsw

Theory and Methodology

3 Performance-based assessment of expertise: How to decide if
4 someone is an expert or not

5 James Shanteau ^{a,*}, David J. Weiss ^b, Rickey P. Thomas ^a, Julia C. Pounds ^c

6 ^a Department of Psychology, Kansas State University, Bluemont Hall 492, 1100 Mid-Campus Drive, Manhattan, KS 66506-5302, USA

7 ^b California State University, Los Angeles, USA

8 ^c Federal Aviation Administration, Oklahoma City, USA

9 Abstract

10 The identification of an expert is vital to any study or application involving expertise. If external criterion (a “gold
11 standard”) exists, then identification is straightforward: Simply compare people against the standard and select whoever
12 is closest. However, such criteria are seldom available for domains where experts work; that’s why experts are needed in
13 the first place. The purpose here is to explore various methods for identifying experts in the absence of a gold standard.
14 One particularly promising approach (labeled CWS for *Cochran–Weiss–Shanteau*) is explored in detail. We illustrate
15 CWS through reanalyses of three previous studies of experts. In each case, CWS provided new insights into identifying
16 experts. When applied to auditors, CWS correctly detected group differences in expertise. For agricultural judges, CWS
17 revealed subtle distinctions between subspecialties of experts. In personnel selection, CWS showed that irrelevant at-
18 tributes were more informative than relevant attributes. We believe CWS provides a valuable tool for identification and
19 evaluation of experts. © 2001 Published by Elsevier Science B.V.

20 *Keywords:* Psychology; Expert systems; Auditing; Management; Agriculture

21 1. Introduction

22 Although experts have been studied for over a
23 century (Shanteau, 1999), there remains a critical
24 unanswered question – how can we describe who
25 is, and who is not, an expert? If there is an external
26 criterion (a “gold standard”), the answer is
27 straightforward. All we have to do is compare a

would-be expert’s judgments to the correct answer. 28
If a person’s answers are close to correct, then he 29
or she is an “expert.” If not, not. 30

This *validity*-based approach is compelling in its 31
simplicity. Unfortunately, it is problematic in ap- 32
plication. The difficulty is that experts are needed 33
precisely in domains where correct answers seldom 34
exist (Gigerenzer et al., 1999; Shanteau, 1995). 35
Indeed, if we could compute (or look up) correct 36
answers, why would we need an expert at all? 37

The purpose of this paper is to explore the ap- 38
plication of a new measure of expertise (labeled 39
CWS for *Cochran–Weiss–Shanteau*) for identifying 40

* Corresponding author. Tel.: +1-785-532-0618; fax: +1-785-532-5401.

E-mail address: shanteau@ksu.edu (J. Shanteau).

41 expertise in the absence of external criteria. The
 42 measure is based on the behavior of would-be
 43 experts by using their performance in the domain.
 44 In effect, this is a bootstrap approach in which the
 45 individual's own decisions are used to validate (or
 46 invalidate) his/her claim to expertise.

47 The remainder of this paper is organized into
 48 five sections: In Section 2, we review approaches
 49 used in prior research to identify would-be experts.
 50 In Section 3, we introduce our proposed approach
 51 to identification of expertise. In Section 4, we ap-
 52 ply this approach to several previously conducted
 53 studies of experts. In Section 5, we consider cave-
 54 ats and restrictions that should be considered
 55 when applying CWS. Finally, we offer our con-
 56 clusions about the future of CWS.

57 2. Prior approaches

58 Many approaches have been used by previous
 59 investigators to identify experts. Nine of these
 60 traditional approaches will be summarized here.
 61 We also consider the advantages and, more im-
 62 portantly, the disadvantages of each approach.

63 2.1. Experience

64 In many studies, the number of years of job-
 65 relevant experience is used as a surrogate for ex-
 66 pertise. Participants with many years of experience
 67 are classified as “experts,” while others with little
 68 experience are labeled “novices.” On the surface,
 69 this approach appears convincing. After all, no
 70 one can function as an “expert” for any length of
 71 time if they are totally incompetent.

72 Although the argument can be made that ex-
 73 perts almost always have considerable experience,
 74 the converse does not necessarily follow. There are
 75 many examples of professionals with considerable
 76 experience who never become experts. Such indi-
 77 viduals may even work with top experts, but they
 78 seldom rise to the performance levels required for
 79 true expertise.

80 In a study of grain judges, for instance, Trumbo
 81 et al. (1962) found that number of years of expe-
 82 rience did not correlate with accuracy of wheat

grading. Instead, their results showed a different 83
 trend: judges with more experience systematically 84
 overrated grain quality (an interesting form of 85
 “grade inflation”). Similarly, Goldberg (1968) 86
 asked clinical psychologists with varying degrees 87
 of experience to diagnose psychiatric patients. He 88
 found no relation between experience and accu- 89
 racy of the diagnoses; however, the confidence of 90
 clinicians in their diagnoses did increase with ex- 91
 perience. 92

Although there are undoubtedly instances where 93
 a positive relationship exists between experience 94
 and performance, there is little reason to expect 95
 this to apply universally. At best, experience is an 96
 uncertain predictor of degree of expertise. At 97
 worst, experience reflects seniority – and little 98
 more. 99

2.2. Certification

100

In many professions, individuals receive some 101
 form of accreditation or title as a reflection of their 102
 skill. For instance, doctors may be certified as 103
 “board certified” and university faculty may be 104
 labeled “full professor.” Generally, it is safe to say 105
 that a certified individual is more likely to be an 106
 expert than someone who is uncertified. 107

The problem with certification is that it is more 108
 often tied to years on the job than it is to profes- 109
 sional performance. This can be particularly true 110
 in bureaucracies. In military photo interpretation, 111
 for instance, the rank of the individuals can vary 112
 from Sergeant to Major. Yet performance is un- 113
 related to rank (Tod Levitt, personal communi- 114
 cation). 115

Another example occurs in the Israeli Air Force, 116
 where the lead pilot in a battle is identified by skill 117
 rather than rank – that means a General may 118
 follow a Captain. This has been cited as one reason 119
 for superiority of the Israelis in air combat against 120
 Arab Air Forces (where lead pilots are usually 121
 determined by rank). The Israelis recognized that 122
 talent is not always reflected by formal certifica- 123
 tion (R. Lipshitz, personal communication). 124

Another problem with certification is the 125
 “ratchet up effect” – people generally move up the 126
 certification ladder, but seldom down. Once certi- 127

128 fied, the recipient is accredited for life. Even if the
 129 skill level of the individual suffers a serious decline,
 130 the title or rank remains. (Just ask students about
 131 the teaching ability of some senior professors.)

132 *2.3. Social acclamation*

133 One method used by many researchers (includ-
 134 ing the present authors) has been to rely on iden-
 135 tification of experts by people working in the field.
 136 That is, professionals are asked whom they con-
 137 sider to be an expert. When there is some agree-
 138 ment about the identification of such an
 139 individual, that person is then labeled an expert by
 140 “social acclamation.”

141 In her analysis of livestock judges, for example,
 142 Phelps (1977) asked professionals in agriculture
 143 whom they considered the best. From their an-
 144 swers, she identified four top livestock judges to be
 145 the experts in her investigation (for further details
 146 on this study, see below).

147 Absent other means of identifying experts, ac-
 148 clamation is a reasonable strategy to follow. It is
 149 unlikely that multiple professional working in a
 150 field would identify the same unqualified person as
 151 an expert. If they agree, it seems safe to assume
 152 that the agreed-upon person is an expert. The
 153 problem with this approach is a “popularity effect”
 154 – someone better known by his or her peers is
 155 more likely to be identified as an expert. Mean-
 156 while, another person outside the peer group is
 157 unlikely to be seen as an expert – even though that
 158 person may be on the cutting edge of new
 159 knowledge. Indeed, those who make new discov-

eries in a field are frequently unpopular in the eyes 160
 of their peers at the time of their breakthroughs. 161

2.4. Consistency (within) reliability 162

Einhorn (1972, 1974) argued that intra-person 163
 (within) reliability is a *necessary condition* for ex- 164
 pertise. That is, an expert’s judgments should be 165
 internally consistent. Conversely, inconsistency 166
 would be *prima facie* evidence that the person is 167
 not an expert. 168

Table 1 lists within-person consistency values 169
 from eight prior studies of experts. The four ver- 170
 tical categories correspond to a classification of 171
 task difficulty proposed by Shanteau (1999). There 172
 are two domains listed for each category, with 173
 internal consistency correlations. For example, the 174
 average consistency for weather forecasters (a de- 175
 cision-aided task) is quite high at 0.98. For 176
 stockbrokers (an unaided task), the average con- 177
 sistency is less than 0.40. 178

As might be expected, aided tasks produce 179
 higher internal consistency values than unaided 180
 tasks. To a first approximation, therefore, it ap- 181
 pears that intra-person reliability corresponds 182
 closely to the performance level of experts in dif- 183
 ferent domains. 184

The difficulty with this approach is that some- 185
 one can be consistent by following some simple, 186
 but incorrect rule. As long as the rule is followed 187
 routinely, the person’s behavior will exhibit high 188
 consistency. For example, by always answering 189
 “yes” and “no” to alternate questions, one can be 190
 perfectly repeatable. But such answers would 191

Table 1
 Reliability (consistency) values across levels of expert performance^a

Highest levels of performance		Lowest levels of performance	
Aided decisions	Competent	Restricted	Unaided decisions
Weather Forecasters <i>r</i> = 0.98	Livestock Judges <i>r</i> = 0.96	Clinical Psychologists <i>r</i> = 0.44	Stockbrokers <i>r</i> < 0.40
Auditors <i>r</i> = 0.90	Grain Inspectors <i>r</i> = 0.62	Pathologists <i>r</i> = 0.50	Polygraphers <i>r</i> = 0.91

^a The values cited in this table (left–right and top–bottom) were drawn from the following: Stewart et al. (1997), Phelps and Shanteau (1978), Goldberg and Werts (1966), Slovic (1969), Kida (1980), Trumbo et al. (1962), Einhorn (1974), and Raskin and Podlesny (1979).

192 generally be inappropriate. Thus, internal consis- 224
 193 tency is a necessary condition – an expert could 225
 194 hardly behave randomly – but not sufficient for 226
 195 expertise. 227

196 *2.5. Consensus (between) reliability*

197 Einhorn (1972, 1974) argued that agreement 228
 198 between individuals is a necessary condition for 229
 199 expertise. That is, he believed that experts in a 230
 200 given field should agree with each other (also see 231
 201 Ashton, 1985). If there is disagreement, then it 232
 202 suggests that one, some, or all of the would-be 233
 203 experts are not really what they claim to be. 234

204 Table 2 lists average between-expert correlations 235
 205 for the same studies listed in Table 1. For instance, 236
 206 the consensus correlations for weather forecasters 237
 207 and stockbrokers are 0.95 and 0.32, respectively. 238
 208 Except for pathologists, the consensus values are 239
 209 similar to, but lower than, the corresponding 240
 210 consistency values in Table 1. 241

211 Livestock judges and polygraphers display quite 242
 212 different consistency and consensus results. Fur- 243
 213 ther analysis reveals that there are several schools 244
 214 of thought in these domains about how to make 245
 215 decisions. Thus, experts from each school may be 246
 216 internally consistent, but show sizable disagree- 247
 217 ment with experts from another school. This could 248
 218 explain the discrepancy between the high consis- 249
 219 tency values and the low consensus values in these 250
 220 two domains. 251

221 On the surface, consensus appears to be a 252
 222 compelling property for experts. After all, we feel 253
 223 quite uncomfortable when two or more experts 254
 255

(such as doctors) argue about which is the correct 224
 procedure to follow. When the experts agree, on 225
 the other hand, we feel more comfortable with the 226
 mutually agreed-upon course of action. 227

The problem with consensus is that agreement 228
 can result from premature closure, e.g., *groupthink* 229
 (Janis, 1972). There are many illustrations where 230
 the best answer was not the one identified by a 231
 group of experts because they focused initially on 232
 an inferior alternative. Thus, they become blind to 233
 better options. Therefore, many experts may agree 234
 – but they may all be wrong (Shanteau, in press; 235
 Weiss and Shanteau, in press). 236

237 *2.6. Discrimination ability*

Hammond (1996) and others have pointed out 238
 that the ability to make fine discriminations be- 239
 tween similar, but not equivalent, cases is a de- 240
 fining skill of experts. That is, an expert must be 241
 able to perceive and act on subtle differences that a 242
 non-expert may often overlook. In the study of 243
 livestock judges by Phelps described below, the 244
 researchers were able to develop quantitative 245
 models of the experts' judgments. However, it 246
 proved impossible for these researchers to apply 247
 the models to actual livestock due to the difficulty 248
 of perceiving the appropriate characteristics of 249
 animals. Thus, knowing *how* to combine infor- 250
 mation is of no value without knowing *what* in- 251
 formation to combine. 252

Although it seems clear that discrimination is a 253
 necessary condition for expertise, there is a catch. 254
 A non-expert may well differentiate between cases 255

Table 2
 Reliability (consensus) values across levels of expert performance^a

Highest levels of performance		Lowest levels of performance	
Aided decisions	Competent	Restricted	Unaided decisions
Weather Forecasters <i>r</i> = 0.95	Livestock Judges <i>r</i> = 0.50	Clinical Psychologists <i>r</i> = 0.40	Stockbrokers <i>r</i> = 0.32
Auditors <i>r</i> = 0.76	Grain Inspectors <i>r</i> = 0.60	Pathologists <i>r</i> = 0.55	Polygraphers <i>r</i> = 0.33

^a The values cited in this table (left–right and top–bottom) were drawn from the following: Stewart et al. (1997), Phelps and Shanteau (1978), Goldberg and Werts (1966), Slovic (1969), Kida (1980), Trumbo et al. (1962), Einhorn (1974), and Lykken (1979).

256 using some easily identifiable, but irrelevant at- 298
 257 tribute. For instance, it is easy to distinguish be- 299
 258 tween livestock based on the length or curliness of 300
 259 their tails. However, tail characteristics play no 301
 260 role in the meat quality of farm animals (Bill Able, 302
 261 personal communication). Thus, discrimination 303
 262 ability is a necessary, but not sufficient, condition 304
 263 for identifying experts. 305

264 2.7. Behavioral characteristics 306

265 Research by (Abdolmohammadi and Shanteau, 307
 266 1992; also see Shanteau, 1989) found that expert 308
 267 auditors share many common behavioral charac- 309
 268 teristics. Some examples are *self-confidence*, *cre-* 310
 269 *ativity*, *perceptiveness*, *communication skills*, and 311
 270 *stress tolerance*. A complete list of characteristics 312
 271 (along with their definitions) appears in the origi- 313
 272 nal paper. 314

273 Because many experts exhibit such traits, Ab- 315
 274 dolmohammadi and Shanteau proposed that be- 316
 275 havioral characteristics might be used to develop a 317
 276 “trait profile” of experts. If appropriate tests can 318
 277 be identified or constructed, then would-be experts 319
 278 would take such tests. Those that score closest to 320
 279 the profile of established experts would then be- 321
 280 come potential experts. 322

281 Although this approach has considerable po- 323
 282 tential, there are three critical problems. First, the 324
 283 required tests for several of these characteristics do 325
 284 not exist, e.g., Communication Skills or Tolerance 326
 285 of Stress. Second, even if they did, the tests would 327
 286 have to be normalized for a domain (e.g., audi- 328
 287 tors). Third, the extent to which non-experts may 329
 288 also share these same characteristics is unclear. 330
 289 Thus, although this approach holds promise, more 331
 290 work is needed before experts can be identified 332
 291 using their behavioral characteristics. 333

292 2.8. Knowledge tests 334

293 In studies of problem solving or game-playing 335
 294 experts are often identified based on tests of fac- 336
 295 tual knowledge. For example, Chi (1978) used 337
 296 knowledge about dinosaurs to separate children 338
 297 into experts and novices. 339

Knowledge of relevant facts is clearly a prereq- 298
 299 uisite for expertise. Someone who knows nothing 300
 301 about a domain will be unable to make competent 302
 303 decisions. Yet, knowledge alone is not sufficient to 304
 305 establish that someone is an expert. In the Chi 306
 307 study, for example, knowledge about different 308
 309 types of dinosaurs is not enough to know what 310
 311 they ate, where they lived, how long they survived, 312
 313 or why they died out. 314

The problem is that it takes more than knowl- 307
 308 edge of facts is needed to be an expert. It is also 309
 310 necessary to see which facts to apply in a given 311
 312 situation. In most domains, that is the hard part. 313

2.9. Creation of experts 311

In certain contexts, it is possible for experts to 312
 313 be “created” through extensive training by re- 314
 315 searchers. This approach has significant advanta- 316
 317 ges, including the fact that development of 318
 319 expertise can be studied longitudinally. Moreover, 320
 321 the skills learned are under direct control of re- 322
 323 searchers. 324

One notable example of this approach is a stu- 319
 320 dent who worked with William Chase at Carnegie- 321
 322 Mellon University to enhance his short-term 323
 324 memory span (Chase and Ericsson, 1981). Because 325
 326 the student was a track athlete, he learned to 327
 328 translate groups of digits into times for various 329
 330 running distances. When asked to retrieve the 331
 332 digits, he recalled the times in clusters tied to 333
 334 running. Using this strategy, the student broke the 335
 336 old record for short-term memory span of 18 digits 337
 338 established by a German mathematician. The new 339
 340 record – over 80! (Other students since have ex- 341
 342 tended the record beyond 100.) 343

Experts can be created in this way for certain 332
 333 narrow tasks, e.g., to play computer games or 334
 335 work in a simulated microworld environment. In 336
 337 most realms of expertise, however, a broad range 338
 339 of skills is required based on years of training and 340
 341 experience. For instance, becoming a medical 342
 343 doctor can take a dozen years just to get started. 344
 345 Obviously, training students for a few months 346
 347 cannot simulate such expertise. 348

341 3. A new approach

342 As the preceding survey shows, many ap- 382
 343 proaches have been advanced for identifying ex- 383
 344 perts. Each of these approaches, however, has one 384
 345 or more serious flaws. No generally acceptable 385
 346 approach exists at the present time. To fill this gap, 386
 347 the two senior authors (Weiss and Shanteau, 387
 348 submitted) proposed a new approach for defining 388
 349 expertise. They combined two necessary, but not 389
 350 sufficient, measures, into a single index. 390

351 First, they agreed with Hammond (1996) that 391
 352 *discrimination* is critical for an expert. The ability 392
 353 to differentiate between similar, but not identical, 393
 354 cases is a hallmark of expertise. That is, experts 394
 355 perceive and act on subtle distinctions that others 395
 356 miss. Second, they followed Einhorn's (1974) 396
 357 suggestion that *consistency*, or within-person reli- 397
 358 ability, is necessary in an expert. If someone can- 398
 359 not repeat their judgment in a similar situation, 399
 360 then they are unlikely to be an expert. 400

361 Discrimination refers to a judge's differential 401
 362 evaluation of different stimulus cases. Consistency 402
 363 refers to a judge's evaluation of the same stimuli 403
 364 over time; inconsistency is its complement. 404
 405

365 3.1. CWS ratio

366 As shown in Eq. (1), Weiss and Shanteau com- 406
 367 bine discrimination and consistency into a ratio. 407
 368 The CWS ratio will be large when a judge dis- 408
 369 criminate consistently, but will be small if the 409
 370 judge either discriminates less or has lower con- 410
 371 sistency. 411

$$372 \text{CWS} = \frac{\text{Discrimination}}{\text{Inconsistency}}. \quad (1)$$

373 Our construction of this index parallels Coch- 412
 374 ran's (1943) suggestion to use a ratio of variances 413
 375 to assess the quality of a response instrument. 414
 376 (Another reason for using variance ratios is that 415
 377 they are asymptotically efficient (I.R. Goodman, 416
 378 personal communication).) Cochran argued that 417
 379 an effective instrument should allow participants 418
 380 to express perceived differences among stimuli in a 419
 381 consistent way. We view an effective expert in the 420

382 same way. We acknowledge our intellectual debt 383
 384 to Cochran by referring to our performance-based 385
 385 index as CWS. 386

387 The intuition underlying the index is that a good 388
 388 measuring tool necessarily has a high CWS ratio. 389
 389 That is, a proper instrument yields different mea- 390
 390 sures for different objects, and gives the same 391
 391 measure whenever it is applied to the same object. 392
 392 A ruler, for example, discriminates among objects 393
 393 of varying length, and produces identical scores 394
 394 for the same objects. Thus, a proper measuring 395
 395 instrument will produce a high CWS value as de- 396
 396 fined in Eq. (1). 397

398 Similarly, an expert must be both discriminating 399
 399 and consistent. It is easy to display one or the 400
 400 other, but hard to do both. One can show dis- 401
 401 crimination by generating a wide variety of re- 402
 402 sponses over stimuli; one can exhibit consistency 403
 403 by repeating the same response to all stimuli. But 404
 404 adopting either of these strategies alone means 405
 405 that the other entity will be lost. To display both 406
 406 properties simultaneously requires careful assess- 407
 407 ment of the stimuli, the essence of expert judg- 408
 408 ment. 409
 409

3.2. Using CWS

410 CWS can be estimated by asking would-be ex- 411
 411 perts to make judgments of a series of stimulus 412
 412 cases; this allows for assessment of their discrimi- 413
 413 nation ability. In addition, at least some of the 414
 414 cases should be repeated; this allows for assess- 415
 415 ment of their consistency. 416

417 Discrimination and inconsistency values can be 418
 418 estimated using a variety of analytic procedures, 419
 419 such as analysis of variance or multiple regression. 420
 420 It is important to emphasize that the use of ratios 421
 421 is descriptive, not inferential. That is, CWS is more 422
 422 of a qualitative tool than a quantitative tool. There 423
 423 are no comparisons to statistical tables and no 424
 424 determinations of significance. Rather, CWS is 425
 425 used to establish that someone behaves more (high 426
 426 value) or less (low value) like an expert. 427

428 To rank-order two (or more) would-be experts, 429
 429 CWS ratios can be compared using a procedure 430
 430 developed by Schumann and Bradley (1959). This 431
 431 allows the researcher to determine whether one 432

427 individual is performing better than another
428 (Weiss, 1985).

429 **4. Reanalyses of prior studies**

430 In this section, we apply CWS to three previous
431 studies of experts. By reanalyzing these results, we
432 hope to show the utility of CWS in a variety of
433 contexts.

434 *4.1. Audit judgment*

435 Ettenson (1984) asked two groups of auditors to
436 evaluate 24 financial cases described by a common
437 set of cues. One group of 15 expert auditors was
438 recruited from Big Six accounting firms in Omaha,
439 Nebraska. The expert group included audit seniors
440 and partners, with 4–25 years of audit experience.
441 For comparison, 15 novice accounting students
442 were obtained from two large Midwestern uni-
443 versities.

444 Every financial case was described using 16 cues,
445 each of which was given either a high or low value.
446 For example, *net income* was set at either a high or
447 low number. For each case, participants were
448 asked to make a *going concern* assessment. A
449 fractional factorial design was used to generate 16
450 cases. Eight of these cases were then replicated to
451 produce a total of 24 stimuli; participants were not
452 told that some cases were identical. The order of
453 presentation of cases was randomized.

454 Based on feedback from an auditor collabora-
455 tor, the cues were classified as either “diagnostic”
456 (e.g., *net income*), “partially diagnostic” (e.g., *aging*
457 *of receivables*), or “non-diagnostic” (e.g., *prior*
458 *audit results*). From analysis of the fractional de-

sign, discrimination was estimated from the mean 459
square values for each cue – high variance implies 460
high discrimination. Inconsistency was estimated 461
from the average of within-cell variances – low 462
variance implies high consistency. The ratio of 463
discrimination variance divided by inconsistency 464
variance was computed to form separate CWS 465
values for diagnostic, partially diagnostic, and 466
non-diagnostic cues. 467

The results in Table 3 show that average CWS 468
values decline systematically as the diagnosticity of 469
the cues declines. For the expert group (first row in 470
Table 3), the differences are notable, especially 471
between diagnostic and partially diagnostic cues. 472
For the novice group (second row in the table), 473
there is a similar but less pronounced decline. 474
More important, there is a sizable difference be- 475
tween experts and novices for diagnostic cues. The 476
size of this difference is less for partially diagnostic 477
cues, and non-existent for non-diagnostic cues. 478

For diagnostic cues, CWS clearly distinguishes 479
between experts and novices. Moreover, the size of 480
difference between the groups declines for less di- 481
agnostic cues. These results show that CWS can 482
distinguish between expert and novice groups. 483

484 *4.2. Livestock judgment*

Phelps (1977) had four professional livestock 485
judges evaluate 27 drawings of gilts – female pigs. 486
These drawings were created by an artist to yield a 487
 $3 \times 3 \times 3$, size \times breeding \times meat quality, factorial 488
design. The judges independently evaluated each 489
gilt for *breeding quality* (how good is the animal 490
for reproduction) and *slaughter quality* (how good 491
is the meat from the animal.) All stimuli were 492
presented three times, although judges were not 493
told that they were being shown the same draw- 494
ings. 495

Two of the judges were nationally recognized 496
experts in assessment of swine and were very fa- 497
miliar with gilts of the sort shown in the drawings. 498
The other two were nationally recognized experts 499
as cattle judges; although they were knowledgeable 500
about swine judging, they lacked day-to-day fa- 501
miliarity and experience. 502

Table 3
Average CWS values for two groups of auditors with three
categories of cues^a

	Diagnostic	Partially diagnostic	Non-diagnostic
Experts	13.10	6.42	3.32
Novices	8.08	5.13	3.03

^a Results based on a reanalysis of Ettenson (1984).

503 For breeding judgments (upper panel in Table
504 4), swine experts produced the largest CWS values
505 for breeding and meat cues. In comparison, cattle
506 experts produced large CWS values only for the
507 meat cue. This apparently reflects the unfamiliarity
508 of breeding characteristics of swine by cattle jud-
509 ges; meat quality characteristics, however, were
510 readily emphasized by cattle judges.

511 For slaughter judgments (lower panel in Table
512 4), the meat cue dominates for both swine and
513 cattle judges. However, there is over a 2-to-1 dif-
514 ference in the magnitude of CWS for meat between
515 swine and cattle judges. Breeding and size dimen-
516 sions were small for both types of judges.

517 Interestingly, for cattle judges, there is little
518 difference in CWS between breeding and slaughter
519 judgments. For swine judges, however, there is a
520 considerable difference between breeding and
521 slaughter judgments, especially for the breeding
522 cue. Thus, it appears that swine judges are more
523 sensitive to changes in the task. In all, CWS pro-
524 vides a revealing picture of the difference between
525 these two highly skilled types of experts. This
526 study also highlights the role that specific tasks
527 play in expertise.

528 4.3. Personnel hiring

529 Nagy (1981) used summary descriptions of job
530 candidates for the position of computer pro-
531 grammer at a large company in the state of
532 Washington. She asked four professional person-
533 nel selectors (experts) and 20 management stu-
534 dents (novices) to evaluate these candidates. Each
535 candidate was described by legally relevant attri-

536 butes (*recommendations from prior employers* and
537 *amount of job-relevant experience*) and legally ir-
538 relevant attributes (*age, gender, and physical at-*
539 *tractiveness*). Filler information from local phone
540 books was used to supply background informa-
541 tion, such as phone number and home address, on
542 the application summaries.

543 Each participant evaluated 32 applicants (gen-
544 erated from a $2 \times 2 \times 2 \times 2 \times 2$ factorial design)
545 twice. Before the evaluations, participants were
546 reminded about the legal requirements for hiring,
547 i.e., what information should and should not be
548 used. The importance of the five attributes was
549 determined for each participant on a 0–100 nor-
550 malized scale; average CWS values are reported
551 for each group.

552 As can be seen for the relevant attributes (upper
553 panel in Table 5), average CWS values are nearly
554 identical for the two groups. This is not surprising
555 given that participants were told immediately be-
556 fore the study about hiring guidelines. In contrast,
557 CWS values for irrelevant attributes (lower panel)
558 reveal a different pattern. For professionals, CWS
559 approaches zero (as it should). In contrast, CWS
560 values are considerably larger for students. Despite
561 being reminded that age, gender, attractiveness,
562 and gender are not legally allowed, business stu-
563 dents had sizable CWS values for these irrelevant
564 attributes. Certainly, it is not easy to ignore
565 something as obvious as age or gender, although
566 that is what the legal guidelines require. Experts,
567 however, apparently have developed strategies to
568 do just that. Thus, there are tasks where CWS
569 values for irrelevant attributes may be more di-
570 agnostic of expertise than relevant attributes.

Table 4
Average CWS values for swine judgments for two types of livestock experts^a

	Size	Breeding	Meat
<i>Breeding judgments</i>			
Swine experts	15.9	53.8	65.6
Cattle experts	< 1.0	3.4	79.2
<i>Slaughter judgments</i>			
Swine experts	< 1.0	3.2	212.7
Cattle experts	< 1.0	7.5	98.0

^a Results based on a reanalysis of Phelps (1977).

Table 5
Average CWS values for two groups of personnel selectors^a

	Recommendations	Experience	
<i>Relevant attributes</i>			
Professionals	88.25	86.17	
Students	88.81	86.88	
	Age	Attractiveness	Gender
<i>Irrelevant attributes</i>			
Professionals	0.99	1.58	0.00
Students	28.12	25.19	13.32

^a Results based on reanalysis of Nagy (1981).

571 5. Caveats

572 There are five caveats and precautions that de-
573 serve mention. First, the application of CWS to
574 these three prior studies is encouraging as far as it
575 goes. However, more evidence is needed before
576 CWS can be used by itself to identify experts. For
577 now, it is clear that CWS can be used as a useful
578 supplement to other approaches, e.g., social ac-
579 clamation.

580 Second, the stimuli used in these studies were
581 abstractions of real-world problems. Specifically,
582 cases were presented in static (non-changing) en-
583 vironments, with no feedback or dynamic/tempo-
584 ral changes. We are now applying CWS in
585 complex, real-time environments.

586 Third, CWS was applied here to individuals
587 whose results were combined to produce group
588 averages. However, most experts work in teams. If
589 teams are treated as a decision-making unit, then it
590 is possible to apply CWS in the same way as with
591 individuals. Preliminary efforts to apply CWS to
592 team decision making have been encouraging.

593 Fourth, CWS assumes that there are real dif-
594 ferences in the stimuli to be judged. If the stimuli
595 are not different, then there is nothing to discrim-
596 inate. If multiple patients have the same disease,
597 for instance, then there will be no differential di-
598 agnoses. Therefore, there must be a range of
599 stimuli before CWS can be used to identify ex-
600 perts.

601 Finally, it is possible for CWS to yield high
602 values for non-experts who use a consistent, but
603 incorrect rule. Suppose all job candidates with
604 short names (e.g., *Ann*) get high recommendations

while all job candidates with long names (e.g., 605
Georgette) get low recommendations. Because of 606
high consistency, such an inappropriate rule would 607
produce high CWS values. One way around this 608
“catch” is to ask judges to evaluate the same cases 609
in different contexts, e.g., recommendations for a 610
different job. If judgments are the same as before, 611
then the participant is not likely to be an expert – 612
despite having a high CWS value. 613

6. Conclusions 614

The present application of CWS leads to five 615
conclusions: First, in the analyses above, CWS 616
proved superior to any previously proposed ap- 617
proach for identifying experts. If CWS continues 618
to be successful, it may provide an answer to the 619
longstanding question of how to identify expertise 620
in the absence of external criterion. 621

622 Second, the success of CWS across different
623 domains is noteworthy. In addition to auditing,
624 livestock judging, and personnel selection, we have
625 applied CWS to wine judging, medical decision
626 making, soil judging, microworld simulations,
627 sensory food evaluations, and air traffic control.
628 Thus far, CWS has worked well in every domain.

629 Third, in addition to identifying experts, CWS
630 has provided new insights into interpretation of
631 previous research. In the Phelps study of livestock
632 judges, for example, CWS clarified a long-standing
633 question about how to distinguish between experts
634 from closely related specialty areas.

635 Fourth, by focusing on discrimination and
636 consistency, CWS may have important implica-

637 tions for selection and training of novices to be-
638 come experts. It is unclear, for example, whether
639 discrimination and consistency can be learned, or
640 whether novices should be preselected for these
641 skills. Either way, CWS offers new perspectives on
642 what it means to be an expert.

643 Finally, we are now applying CWS to data sets
644 where there is no prior information about the
645 relevance of attributes. The question is whether
646 CWS can identify experts in the absence of any
647 knowledge of what is relevant and what is irrele-
648 vant. In preliminary analyses, the differences do
649 not appear to be as large as shown in the present
650 tables. However, CWS does consistently separate
651 experts from non-experts. In all, the future for
652 CWS looks hopeful.

653 Acknowledgements

654 Preparation of this manuscript was supported,
655 in part, by grant 96-12126 from the *National Sci-*
656 *ence Foundation* and by grant 98-G-026 from the
657 *Federal Aviation Administration* in the *Department*
658 *of Transportation* (in the USA).

659 References

- 660 Abdolmohammadi, M.J., Shanteau, J., 1992. Personal charac-
661 teristics of expert auditors. *Organizational Behavior and*
662 *Human Decision Processes* 58, 158–172.
- 663 Ashton, A.H., 1985. Does consensus imply accuracy in
664 accounting studies of decision making. *Accounting Review*
665 60, 173–185.
- 666 Chase, W.G., Ericsson, K.A., 1981. Skilled memory. In:
667 Anderson, J.R. (Ed.), *Cognitive Skills and Their Acquisi-*
668 *tion*. Erlbaum Associates, Hillsdale, NJ, pp. 141–189.
- 669 Chi, M.T.H., 1978. Knowledge structures and memory devel-
670 opment. In: Siegler, R.S. (Ed.), *Children's Thinking: What*
671 *Develops?* Erlbaum Associates, Hillsdale, NJ, pp. 73–96.
- 672 Cochran, W.G., 1943. The comparison of different scales of
673 measurement for experimental results. *Annals of Mathe-*
674 *matical Statistics* 14, 205–216.
- 675 Einhorn, H.J., 1972. Expert measurement and mechanical
676 combination. *Organizational Behavior and Human Per-*
677 *formance* 7, 86–106.
- 678 Einhorn, H.J., 1974. Expert judgment: Some necessary condi-
679 tions and an example. *Journal of Applied Psychology* 59,
680 562–571.

- Ettenson, R., 1984. A schematic approach to the examination 681
of the search for and use of information in expert decision 682
making. Unpublished Doctoral Dissertation, Kansas State 683
University, Manhattan, KS. 684
- Gigerenzer, G., Todd, P., & the ABC group, 1999. *Simple* 685
Heuristics that Make Us Smart. Oxford University Press, 686
London. 687
- Goldberg, L.R., 1968. Simple models or simple processes Some 688
research on clinical judgments. *American Psychologist* 23, 689
482–496. 690
- Goldberg, L.R., Werts, C.E., 1966. The reliability of clinicians 691
judgments: A multitrait-multimethod approach. *Journal* 692
of Consulting Psychology 30, 199–206. 693
- Hammond, K.R., 1996. *Human Judgment and Social Policy*. 694
Oxford University Press, New York. 695
- Janis, I.L., 1972. *Victims of Groupthink*. Houghton-Mifflin, 696
Boston. 697
- Kida, T., 1980. An investigation into auditor's continuity and 698
related qualification judgments. *Journal of Accounting* 699
Research 22, 145–152. 700
- Lykken, D.T., 1979. The detection of deception. *Psychological* 701
Bulletin 80, 47–53. 702
- Nagy, G.F., 1981. How are personnel selection decisions made 703
An analysis of decision strategies in a simulated personnel 704
selection. Unpublished Doctoral Dissertation, Kansas 705
State University, Manhattan, KS. 706
- Phelps, R.H., 1977. Expert livestock judgment: A descriptive 707
analysis of the development of expertise. Unpublished 708
Doctoral Dissertation, Kansas State University, Manhat- 709
tan, KS. 710
- Phelps, R.H., Shanteau, J., 1978. Livestock judges: How much 711
information can an expert use?. *Organizational Behavior* 712
and Human Performance 21, 209–219. 713
- Raskin, D.C., Podlesny, J.A., 1979. Truth and deception: A 714
reply to Lykken. *Psychological Bulletin* 86, 54–59. 715
- Schumann, D.E.W., Bradley, R.A., 1959. The comparison of 716
the sensitivities of similar experiments: Model II of the 717
analysis of variance. *Biometrics* 15, 405–416. 718
- Shanteau, J., 1989. Psychological characteristics and strategies 719
of expert decision makers. In: Rohrman, B., Beach, L.R., 720
Vlek, C., Watson, S.R. (Eds.), *Advances in Decision* 721
Research. North-Holland, Amsterdam, pp. 203–215. 722
- Shanteau, J., 1995. Expert judgment and financial decision 723
making. In: Green, B. (Ed.), *Risky Business*, University of 724
Stockholm School of Business, Stockholm, pp. 16–32. 725
- Shanteau, J., 1999. Decision making by experts: The GNAHM 726
effect. In: Shanteau, J., Mellers, B.A., Schum, D.A. (Eds.), 727
Decision Science and Technology: Reflections on the 728
Contributions of Ward Edwards. Kluwer Academic Pub- 729
lishers, Boston, pp. 105–130. 730
- Shanteau, J., in press. What does it mean when experts 731
disagree? In: Salas, E., Klein, G. (Ed.), *Linking Expertise* 732
and Naturalistic Decision Making. Erlbaum Associates, 733
Hillsdale, NJ. 734
- Slovic, P., 1969. Analyzing the expert judge: A descriptive study 735
of a stockbroker's decision processes. *Journal of Applied* 736
Psychology 53, 255–263. 737

- 738 Stewart, T.R., Roebber, P.J., Bosart, L.F., 1997. The impor- 745
739 tance of the task in analyzing expert judgment. *Organiza- 746*
740 tional Behavior and Human Decision Processes 69, 205– 747
741 219. 748
- 742 Trumbo, D., Adams, C., Milner, M., Schipper, L., 1962. 749
743 Reliability and accuracy in the inspection of hard red 750
744 winter wheat, *Cereal Science Today* 7. 751
- Weiss, D.J., 1985. SCHUBRAD: The comparison of the 745
sensitivities of similar experiments. *Behavior Research 746*
Methods Instrumentation and Computers 17, 572. 747
- Weiss, D.J., Shanteau, J., in press. The Vice of Consensus and 748
the Virtue of Consistency. In: Shanteau, J., Johnson, P., 749
Smith, C. (Eds.), *Psychological Explorations of Competent 750*
Decision Making. Cambridge University Press, New York. 751
- Weiss, D.J., Shanteau, J., submitted. Empirical assessment of 752
expertise. 753