Journal of Applied Psychology
2013, Vol. 98, No. 6, 1060–1072

0021-9010/13/$12.00 DOI: 10.1037/a0034156

RESEARCH REPORT

Mechanical Versus Clinical Data Combination in Selection and Admissions Decisions: A Meta-Analysis

Nathan R. Kuncel
University of Minnesota

David M. Klieger
Educational Testing Service, Princeton, New Jersey

Brian S. Connelly
University of Toronto

Deniz S. Ones
University of Minnesota

In employee selection and academic admission decisions, holistic (clinical) data combination methods continue to be relied upon and preferred by practitioners in our field. This meta-analysis examined and compared the relative predictive power of mechanical methods versus holistic methods in predicting multiple work (advancement, supervisory ratings of performance, and training performance) and academic (grade point average) criteria. There was consistent and substantial loss of validity when data were combined holistically—even by experts who are knowledgeable about the jobs and organizations in question—across multiple criteria in work and academic settings. In predicting job performance, the difference between the validity of mechanical and holistic data combination methods translated into an improvement in prediction of more than 50%. Implications for evidence-based practice are discussed.

Keywords: judgment and decision making, mechanical versus clinical data combination, criterion related validity

Predicting performance in work and academic settings is quite complex, with a large utility for strong prediction. Since numerous individual and situational factors have been shown to influence performance and both jobs and some performance determinants can change over time, multiple measures are often used to thoroughly evaluate applicants. For even moderately complex jobs, a great deal of information frequently is collected via tests, interviews, resumes, and simulations, creating the ultimate issue of how to best make use of it all.

Two general approaches have been used to combine data collected from applicants. The first are mechanical (actuarial, algorithmic) approaches that involve applying an algorithm or formula to each applicant's scores. Examples range from aggregating scores using simple unit weights, to estimating optimal weights, to using more complex empirically derived decision trees. Holistic methods, the second general and more common approach (clinical, expert judgment, intuitive, subjective), include both individual judgments of data and group consensus meetings. The defining characteristic of the holistic methods is that data are combined using judgment, insight, or intuition, rather than an algorithm or formula that is applied the same way for each decision.

Although the holistic approach has remained the most common approach over time (Jeanneret & Silzer, 1998; Ryan & Sackett, 1987), previous research across a range of fields has demonstrated consistently improved decision accuracy for mechanical methods over holistic ones (Grove & Meehl, 1996). Several reviews have been conducted evaluating and comparing different types of mechanical and clinical data combination across a mixture of fields and decision types (e.g., Grove, Zald, Lebow, Snitz, & Nelson, 2000; Sawyer, 1966). There are two consistent findings in these reviews. The first is that the specific type of mechanical versus holistic method is largely less important than whether or not the method uses human judgment versus an equation for data combination. Second, the central issue appears to be how the data are combined together to form a judgment or recommendation rather than how they are gathered in the first place. That is, people are effective at collecting information but appear to be less effective at combining multiple sources of information into a final decision.

However, no meta-analysis has been conducted on this issue for the prediction of human performance in work and academic settings. Such an investigation is important because the actual size of the difference between mechanical and holistic approaches in predicting work or academic performance is unknown. Given that there is generally a strong preference for holistic expert-driven clinical decision making (Highhouse, 2008b) in Industrial-Work

This article was published Online First September 16, 2013.

Nathan R. Kuncel, Department of Psychology, University of Minnesota; David M. Klieger, Educational Testing Service, Princeton, New Jersey; Brian S. Connelly, Department of Management, University of Toronto, Toronto, Ontario, Canada; Deniz S. Ones, Department of Psychology, University of Minnesota.

Correspondence concerning this article should be addressed to Nathan R. Kuncel, Department of Psychology, University of Minnesota, 75 River Road, Minneapolis, MN 55455-0344. E-mail: kunce001@umn.edu

and Organizational (IWO) psychology, a relatively small difference would suggest that the method of data combination is a marginal issue. Consequently, emphasis should be placed on encouraging the consistent use and ongoing development of high-quality predictors. On the other hand, if the difference is large, then research is needed to understand its source and find methods that capture at least some of the strengths of the mechanical methods while remaining acceptable to end users.

Brunswik Lens Model

To frame the current study, we adopt the Brunswik Lens Model (Brunswik, 1955, 1956), which provides a theory of decision making and an elegant analytical framework. Conceptually, the Lens Model assumes that people perceive information in their environment and combine one or more pieces of information into a judgment or prediction. In a selection context, this information could be anything from characteristics of the setting to subtle behavior cues from potential job candidates to an understanding of the foibles of senior management. The human judge can weight each piece differentially and then combine the information to yield a prediction or judgment. The Lens Model permits modeling the human judge's weighting and combining of information cues with any combination of methods (additive, configural, power, interactive, conditionally) and comparing it to other methods of data combination.

Structurally, the Lens Model contains three major components: the subject response or judgments (Y_s), the environmental or independent variable cues (information cues), and the outcome or criterion value of interest (Y_e). The relations among these components is used to evaluate the nature of decision making (see Figure 1). Specifically, the Lens Model specifies that the judgment made by an individual is based on their perception and mental weighting of one or more cues (e.g., observed interview behavior, test scores, resume items). These cues, in turn, have actual associations with an outcome (e.g., performance, turnover). One can think of the judge peering into the future through the lens of the environmental cues influenced by the weights and combinations used by the judge.

Typically multiple regression is used to quantify how cues are related to judgments as well as outcomes. Regressing the judgment on the cues (called the Cognitive Strategy) and the outcome on the cues (called the Ecological Validity) models how the cues, on average, are related to judgments and outcomes, respectively. Correlating the judgments with the outcome is often called the Achievement Index and estimates how strongly judgments are predictive of outcomes. This meta-analysis contrasts the magnitude of the Ecological Validity with the Achievement Index. However, much more can be done with the model and previous research on other aspects of the model can aid in interpreting the current results. The Lens Model is particularly powerful because it allows scholars to examine how cues are typically used but also how variably.

It is well established that judges use cues "inconsistently" in that they deviate in many cases from the estimated regression weights based on the judge's predictions (e.g., Karalaia & Hogarth, 2008). In other words, judges will often weight the same set of cues differently across targets, weighting, for example, historical accomplishments more for one candidate than another. Thus, the model makes a distinction between "man" (the judge's prediction for each individual target) and the "model of man" (the estimated average values from regressing the judge's prediction on a set of cues).

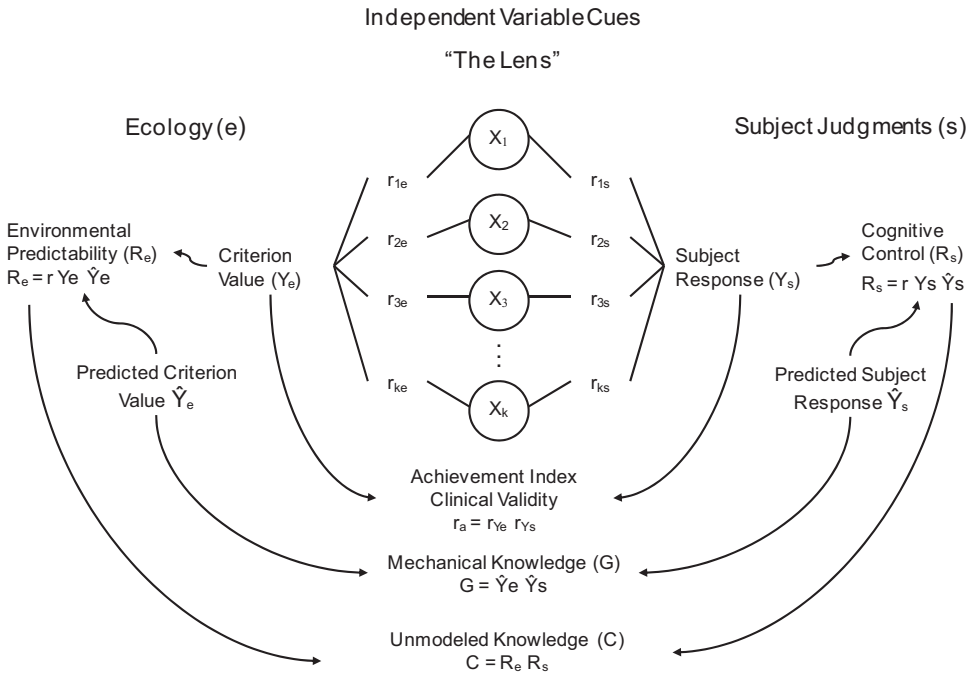

Figure 1. The Lens Model.

The correspondence between the "man" and the "model of man" predictions has been referred to as "cognitive control" (R_s; Hammond & Summers, 1972). It quantifies just how consistently judges combine information. Cognitive control is often fairly low for judges making complex psychological judgments. This is not problematic, per se, if the deviations made by the judge improve judgment accuracy. For example, intuiting that biodata are especially salient for one applicant based on structured interview results would lead to lower cognitive control (i.e., biodata are not consistently weighted) but might yield improved prediction accuracy.

This leads to one of the most striking findings from Lens Model research. Models of the judge consistently outperform the judge's actual judgments. Remarkably, adhering to a weighted composite based on previous judgments will do better than the expert on whom it was developed (e.g., Goldberg, 1970). Returning to our example, this evidence suggests that the judge's intuition to more heavily emphasize the biodata information for one individual will typically be in error and that consistent use of the cues tends to results in better predictions.

Theoretically, judge inconsistency leaves open the possibility that the judgments do, in fact, contain insights that improve on a mechanical composite but the concern is that these insights may be plagued by unreliability. The Lens Model provides two mechanisms for evaluating this question. First, the judge's predictive validity above and beyond a linear optimal weighting of cues can be examined. This is called "C" in Lens Model parlance (or Unmodeled Knowledge) and is the correlation of the residuals between the Cognitive Strategy model and the Ecological Validity model. C is in contrast to G, which is called Mechanical Knowledge. Mechanical knowledge is the correspondence between predictions made by an optimally weighted mechanical combination of predictors and predictions made the human judge. These different indexes allow scholars to examine the extent to which judges employ less than optimal weighting schemes.

Research has indicated that there is little predictive power unique to clinical judgments (Karalaia & Hogarth, 2008). On the whole, these findings suggest that, compared to mechanical prediction from a set of cues, clinical judgment is degraded by *both* failing to appropriately weight all relevant cues (i.e., $R_s < 1$) and unreliably in applying weighting schemes that are developed (i.e., $G < 1$). This degradation does not generally appear to be compensated with unique accuracy of clinical predictions as C values (insights beyond the linear mechanical model) are typically zero or very small (Karalaia & Hogarth, 2008). Although previous theory and research provide a good explanation of the cause of any differences, the impact of inappropriate weights and unreliability in applying those weights on decisions varies by topic. Estimating the size of the difference is a major reason for the present study. Although this introduction has focused the Lens Model on the mechanical/clinical question, it has much broader implications for validation research and we return to the model's implications in the discussion. We argue that it should be adopted as an overarching framework for all personnel selection research.

In total, theory and prior research suggest that an over attention to salient cues and inconsistency in use of weights without a compensating gain in insights will cause experts to form judgments with less predictive power than mechanical combination of the same set of cues. The questions remain, "How much of a difference?" and "For what criteria?" This meta-analysis examines and compares the relative predictive power of mechanical methods versus clinical methods in predicting multiple work (advancement, supervisory ratings of performance, and training performance) and academic (grade point average) criteria.

Method

The Meta-Analytic Database

We used a modified version of Hunter and Schmidt's (2004) psychometric meta-analytic method (see below) to quantitatively aggregate results across studies that compared criterion-related validities of clinical data combination methods to criterion-related validities of mechanical data combination methods. Studies were gathered from several sources. Using the terms "combination," "mechanical," "actuarial," "clinical," "impressionistic," "holistic," "fit," and "judgmental" (as well as synonyms and different forms of these terms) to identify relevant research, we searched PsycINFO (1887–2008), ERIC (Education Research Information Center, 1966–2008), Digital Dissertations (1861–2008), and google.com (2008). We examined citation lists within all articles, dissertations, and technical reports to obtain additional relevant studies.

To be included in the database, a study had to quantitatively compare use of mechanical combination of data from one or more independent variables to the use of clinical combination of the same data from the same independent variables to predict work or academic criteria (e.g., performance, achievement).[1] Thus, each study included (a) at least one effect size for a mechanical data combination method correlated with a work or academic criterion and (b) at least one effect size for a clinical data combination method correlated with the same exact criterion.

The independent variables used for each data combination method were selected to maximize their similarity. That is, to the extent possible we made "apples versus apples" comparisons where the clinician used and had access to the same information as was used in the mechanical combination. In no case could the mechanical combination methods use information that was unavailable to the clinician, as such comparisons would favor the mechanical method. Some studies compared clinical and mechanical methods for varying numbers of predictors, including scenarios in which one method had more predictors than the other. However, for each such study we also chose the closest match in the number of predictors for each combination method (provided that the mechanical method did not employ more predictors while allowing the clinical some leeway). The only exception was when both methods had technically different measures that effectively measured the same construct and are very similar in predictive power (e.g., high school rank vs. high school grade point average; Neidich, 1968).

[1] The one partial exception to this rule is the comparison of the mechanical data combination of meta-analyzed dimension scores reported in Arthur, Day, McNelly, and Edens (2003; $R = .45$), which Arthur et al. compared to the meta-analyzed overall assessment rating (OAR) reported in Gaugler, Rosenthal, Thornton, and Bentson (1987; corrected $r = .37$). It is not fully clear the extent to which the clinicians whose clinical data combination methods reflected in Gaugler et al.'s correlation of .37 had access to the same information used in the mechanical data combination procedures of Arthur et al.

Furthermore, we did not use effect sizes for the mechanical data combination method that were the result of exploratory analytic techniques that would gave it an unfair advantage over the clinical data combination method. For example, in some studies (e.g., Lewis & MacKinney, 1961; Mitchel, 1975), authors selectively chose a subset of independent variables for the mechanical combination only after screening based on relationships with the criterion. Such selectivity may capitalize on sampling error and overestimate true predictive power of the mechanical combination.

Coding of all articles was inspected by two or more authors. To avoid violating assumptions of the independence of samples, effect sizes were first averaged within a particular study prior to averaging across studies. Two studies (Arthur, Day, McNelly, & Edens, 2003; Stuit, 1947) had relatively much larger sample sizes (*N*). To prevent these studies from overwhelming any analysis, effect sizes from these studies were weighted by the median of other studies' sample sizes rather than their own sample size. However, it should be noted that inclusion of these studies with the full *N*s does not alter the conclusions of the study.[2]

Meta-Analytic Procedures

Selection of workers and students on the basis of predictors often results in range restriction, which in turn attenuates estimates of predictive validity. Unreliability of predictor and criterion measures also attenuates these estimates. Unfortunately, there were inadequate sample specific data to correct studies for either range restriction or unreliability either individually or through artifact distributions. Hence, the validity estimates provided here are likely to be underestimates of the actual relationship between predictors and criteria. Not correcting for statistical artifacts means that the magnitudes of differences between mechanical and clinical combination methods may also be underestimated. For example, if the criterion reliability is $r_{yy} = .60$, comparing an observed validity of .25 to another observed validity of .30 underestimates the difference in their corrected correlations (.32 vs. .39).

For all clinical data combination, effect sizes were obtained (either directly or through calculation) as zero-order correlations. In the case of mechanical combination, effect sizes were either zero-order correlations (*r*s) or multiple correlations (multiple-*R*s). However, these multiple correlations are upwardly biased because predictions made with more than one predictor may capitalize on chance factors specific to a particular sample for which regression weights are estimated (Nunnally, 1978). This capitalization on chance results in a multiple correlation value that will typically overestimate the mechanical formula's predictive validity in another sample or the population (i.e., shrinkage). The research purpose here is to ascertain how well the independent variables predict in future applied settings. The two estimates that can be created are the estimated cross-validated multiple correlation and the population multiple correlation.

Given that most selection systems are ongoing and weights can be refined over time, neither estimator is ideal. The cross-validated multiple correlation can be considered as the value of the regression weights for the subsequent set of decisions. (ρ_c; Cattin, 1980a; Fowler, 1986). ρ_c (or its estimate $\hat{\rho}_c$) indicates how predictive a formula is when the regression weights are created based on data from one sample and then reused in subsequent samples drawn from the same population.[3] However, with efforts to refine the weights with more data, the estimated cross-validated multiple correlation would then be the lower bound with the estimated population Multiple-R providing as estimate of the upper bound. Therefore, we provide results based on both sets of estimates. To estimate $\hat{\rho}_c$, we used a version of Browne's (1975) formula (see the Appendix, Formula 3). Although alternate methods of calculating $\hat{\rho}_c$ exist (e.g., Claudy, 1978; Rozeboom, 1978), this version of Browne's formula has generally performed best in Monte Carlo simulation studies (Raju, Bilgic, Edwards, & Fleer, 1999; Shieh, 2008; Yin & Fan, 2001).

An adapted version of Hunter and Schmidt's (2004) "bare-bones" meta-analytic procedure was used to aggregate results across studies to estimate the mean predictive validity, the observed variability around that mean predictive validity, and the variability remaining after accounting for variability due to sampling error. The modification was necessary due to the mixture of effect sizes included in the mechanical estimates. In estimating the variability due to sampling error, our combination of zero-order correlations and multiple-R correlations for mechanical combinations necessitated adapting Hunter and Schmidt's bare-bones procedures. Specifically, although sampling error impacts our estimate of a population-level effect size whether the sample-level effect sizes in our meta-analyzed studies are zero-order correlations or multiple correlations, the formulae for estimating sampling error of these statistics differ.

Therefore, we estimated sampling error variance individually for each sample using the appropriate sampling error statistic for the effect size (r or $\hat{\rho}_c$). For each of the zero-order correlations, σ_e^2 (sampling error variance) was calculated using Hunter and Schmidt's (2004, pp. 85–92) bare-bones procedure. For each of the multiple correlations, a measure of variability for each effect size point estimate, $var(\hat{\rho}_c)$—which Browne (1975) refers to as $var(\omega)$—was calculated using Browne's estimation method (see the Appendix, Formula 4).[4]

To estimate true variability around the meta-analytic mean, the individual sample estimates of σ_e^2 and $var(\omega)$ were pooled together. An observed sample-size weighted correlation variance was calculated using the mean observed effect size, zero-order correlations, and the shrunken *R*s. Hunter and Schmidt's (2004, pp. 85–92) bare-bones procedure provides the appropriate formula with an example. The pooled error variance was subtracted from the observed correlation variance, and then the square root was taken to obtain SD_ρ.

The final database included 25 samples across 17 studies. After replacing extreme sample outliers with the median of the *N*s for the other samples in the same analysis, there were 2,263 workers for whom predictions were made via mechanical data combination, 2,027 workers for whom predictions were made via clinical data

[2] For interested readers, these results are available from the first author.

[3] $\hat{\rho}_c$ is preferable to alternate formulas for adjusting for shrinkage, such as Cattin's (1980a) ρ, because the goal of these analyses is to estimate validities that would be observed in a sample (a new set of applicants). Cattin's ρ, however, is appropriate when the goal is to estimate the population-level multiple correlation.

[4] Although several competing approaches exist for estimating $var(\omega)$ (Fowler, 1986; Mendoza & Stafford, 2001), these approaches involve added complexity without demonstrated improvement over Browne's (1975) method. The use of Browne's formulas was most with a preference for transparency and parsimony in methodology.

combination, 889 students for whom predictions were made via mechanical data combination, and 632 students for whom predictions were made via clinical data combination. Within each analysis, the samples were independent of each other and the effect sizes included relationships for three work and two academic criteria.

Results

Summaries of each study contributing to the meta-analysis are presented in Table 1. For those estimates that required aggregation of Multiple *R*s, the magnitude of the mechanical estimate varied depending on the shrinkage formulae applied. When aggregated separately for each outcome variable, across all outcome variables for population estimates, a consistent pattern emerged. Larger correlations were found for mechanical methods over clinical methods. For many important criteria, validities of mechanical methods were substantially larger than those found for clinical data combination approaches (see Table 2). The mechanical advantage was eliminated but not reversed for two criteria when the most stringent new-sample shrinkage estimates were employed.

For job performance, the average correlation was .44 for mechanical and .28 for clinical. Advancement criteria yielded a smaller difference of .42 for mechanical versus .36 for clinical. The least data existed for training outcomes but the results were consistent with other results with an average of .31 for mechanical and .16 for clinical. For the educational criterion of grade point average, the predictive validities were larger with an average value of .58 for mechanical and .48 for clinical. Finally, in our most diverse analysis, a collection of three different measures of nongrade measures of academic achievement (faculty evaluations, comprehensive exam performance, and degree completion) yielded the narrowest difference with an average of .47 for mechanical and .46 for clinical prediction (although the latter is based on only 161 students).

Most of these differences in validity are substantial, especially for job performance. In predicting this criterion, the difference between the validity of mechanical and clinical data combination methods translates into a population level improvement in prediction of more than 50%.

Discussion

The results of this meta-analysis demonstrate a sizable predictive validity difference between mechanical and clinical data combination methods in employee selection and admission decision making. For predicting job performance, mechanical approaches substantially outperform clinical combination methods. In Lens Model language, the Achievement Index (clinical validity) is substantially lower than the Ecological Validity.

This finding is particularly striking because in the studies included, experts were familiar with the job and organizations in question and had access to extensive information about applicants. Further, in many cases, the expert had access to more information about the applicant than was included in the mechanical combination. Yet, the lower predictive validity of clinical combination can result in a 25% reduction of correct hiring decisions across base rates for a moderately selective hiring scenario (SR = .30; Taylor & Russell, 1939). That is, the contribution our selection systems make to the organization in increasing the rate of acceptable hires is reduced by a quarter when holistic data combination methods are used. Yet, this is an underestimate because we were unable to correct for measurement error in criteria or range restriction. Corrections for these artifacts would only serve to increase the magnitude of the difference between the methods.

Despite the results obtained here, it might be argued that one great advantage of the clinical method is that frequent changes in jobs or circumstances will lead to a situation where the equation is no longer appropriate while a clinical assessment can accommodate the change in circumstances. There are three problems with this argument. First, there is no empirical evidence supporting this scenario in the literature. The performance dimensions of jobs have remained quite stable over time. For example, early evaluations of the dimensional structure of the job of managers yielded much the same dimensions as contemporary models (e.g., Borman & Brush, 1993; Campbell, Dunnette, Lawler, & Weick, 1970; Flanagan, 1951). Second, linear models are quite robust to changes in weights. That is, unless the weights suddenly change from positive to negative (another situation that has never been observed in the literature), the overall predictive power of the composite remains strong (Dawes, 1979). Finally, if such a situation were to occur, the use of an expert's subjective weights, integrated into a modified equation, would still outperform the clinician.

Small *N* situations are also sometimes raised as a concern. It is sometimes argued that these settings prevent the use of mechanical methods. This is not the case. Each predictor can be weighted by evidence from the literature (e.g., dominance is typically a moderately valid predictor of leadership effectiveness). The advent of validity generalization and considerable number of meta-analyses in the literature provides some solid ground for differential weighting. Alternatively, expert judgment (preferably aggregated across multiple experts) can be used to set weights (e.g., our stock in-basket should get only nominal attention given the job level and functional area). These values can then be used to weight and combine the assessment results.

The field would benefit from additional research that investigates specific, and hopefully controllable, features of the assessment, assessee, and decision process that contribute to reduced predictive power. It is possible that assessors are overly influenced by aspects of candidate's personality or demeanor that are not associated with subsequent job performance. Such evidence could be used for assessor training to reduce such systematic errors and could be combined with methods to increase the use of effective predictors and data combination methods (Kuncel, 2008). Although the results presented here are wholly consistent with a broader literature, ongoing research is important for expanding on the modest number of studies presenting evidence of this comparison. The file drawer problem could also be present although we expect that, given common practice, results would tend to skew in favor of mechanical rather than holistic judgment.

Viewing and Reframing Personnel Selection Through the Lens Model

From our perspective, the Lens Model provides a new way of thinking about personnel selection that reaches well beyond the issue of mechanical versus expert judgment. The true focus in validation work should be on how information is used and what decisions are

Table 1
Studies Contributing to the Meta-Analysis

Analyses in which included	Authors (Year)	Criteria	Predictors	$Type_{mech}$	r_{mech}	N_{mech}	$Type_{clin}$	r_{clin}	N_{clin}
1. Acad.–GPA	Sarbin (1943)	Acad. Ach.	Achievement[1,2]; cognitive ability[1,2]; vocational interest[1,2]; personality[1,2]; records[1,2]; interviews[1,2]	MR	0.70	89	Judgment of clinical counselors	0.69	89
2. Acad.–GPA	Sarbin (1943)	Acad. Ach.	Achievement[1,2]; cognitive ability[1,2]; vocational interest[1,2]; personality[1,2]; records[1,2]; interviews[1,2]	MR	0.45	73	Judgment of clinical counselors	0.35	73
3. Acad.–GPA	Stuit (Ed.) (1947)	Acad. Ach.	Cognitive ability[1,2] and interviews (with predictor scores available to interviewer)[2]	C & r	0.50	3,246 (73)[3]	Interview	0.41	3,246 (89)[3]
4. Acad.–Non-Grade	Truesdell & Bath (1957)	Acad. persistence	Achievement[1,2]; vocational interests (inventory)[2]; vocational interests (subscales)[1]; personality (inventories)[2]; personality (subscales)[1]	DF	0.50	314	Average of validities of academic staff judgments	0.42	100
5. Acad.–GPA	Watley & Vance (1964)	Acad. Ach.	Student's age[2]; name of high school from which the student graduated[2]; achievement[1,2]; plans for academic major[2]; cognitive ability[1,2]	MR	0.61	71	Counselor judgment	0.54	100
6. Acad.–Non-Grade	Robertson & Hall (1964)	Acad. Ach.	Cognitive ability[1,2]; achievement (college)[1,2]; achievement (grade-level)[2]	S-DW	0.54	38	Faculty ratings	0.74	38

(*table continues*)

Table 1 (*continued*)

Analyses in which included	Authors (Year)	Criteria	Predictors	$Type_{mech}$	r_{mech}	N_{mech}	$Type_{clin}$	r_{clin}	N_{clin}
7. Work–Advancement	Wollowick & McNamara (1969)	Job advancement	Personality[1,2]; leadership[1,2]; cognitive ability[1,2]; demographics[1,2]; assessment center exercises[1,2]; and various personality, cognitive, and specific ability "characteristics"[1,2]	MR	0.62	94	Assessment staff ratings	0.37	94
8. Acad.–Non-Grade	Dawes (1971)	Acad. Ach.	Cognitive ability[1,2]; achievement[1,2]; and quality of undergraduate institution[1,2]	MR	0.40	111	Average rating made by the admissions committee	0.19	23
9. Acad.–GPA	Wiggins & Kohen (1971)	Acad. Ach.	Cognitive ability[1,2]; achievement[1,2]; undergraduate school selectivity[1,2]; personality[1,2]; and the sex of the student[1,2]	MR; C-B; and BW	0.59	90	Average predicted judgments of psychology graduates and average validity across judges	0.40	90
10. Acad.–GPA	Nystedt & Magnusson (1972)	Acad. Ach.	Cognitive ability[1,2]; intercorrelations between predictor tests[2]; and predictor tests' ecological validity[2]	MR	0.73	30	Group judgment	0.47	30
11. Work-–Performance	Huck (1974)—White sample	Overall job performance	Interviews[1,2]; assessment center exercises[1,2]; cognitive ability[1,2]; interests[1,2]; written communication[1,2]; and biodata[1,2]	MR	0.625	91	Individual assessors' overall judgments adjustable after group discussion	0.41	91
12. Work–Performance	Huck (1974)—Black sample	Overall job performance	Interviews[1,2]; assessment center exercises[1,2]; cognitive ability[1,2]; interests[1,2]; written communication[1,2]; and biodata[1,2]	MR	0.78	35	Individual assessors' overall judgments adjustable after group discussion	0.35	35

Table 1 (*continued*)

Analyses in which included	Authors (Year)	Criteria	Predictors	$Type_{mech}$	r_{mech}	N_{mech}	$Type_{clin}$	r_{clin}	N_{clin}
13. Work–Advancement	Huck (1974)—White sample	Potential for advancement	Interviews[1,2]; assessment center exercises[1,2]; cognitive ability[1,2]; interests[1,2]; written communication[1,2]; and biodata[1,2]	MR	0.67	91	Individual assessors' overall judgments adjustable after group discussion	0.59	91
14. Work–Advancement	Huck (1974)—Black sample	Potential for advancement	Interviews[1,2]; assessment center exercises[1,2]; cognitive ability[1,2]; interests[1,2]; written communication[1,2]; and biodata[1,2]	MR	0.82	35	Individual assessors' overall judgments adjustable after group discussion	0.54	35
15. Work–Training	Borman (1982)	Job training performance	Interviews (structured)[1,2] and assessment center exercises[1,2]	P-UW	0.39	47	Judgment by assessors	0.30	47
16. Work–Advancement	Tziner & Dolan (1982)	Job training performance	Interviews[1,2]; assessment center exercises[1,2]; supervisor evaluations[1,2]; cognitive ability[1,2]; and personality[1,2]	MR	0.47	193	Group judgment by assessors	0.38	193
17. Work–Performance	Feltham (1988)	Job performance	Assessment center exercises[1,2] and cognitive ability[2]	C-UW	0.28	141	Group judgment by assessors	0.16	141
18. Work–Training	Feltham (1988)	Job training performance	Assessment center exercises[2] (committee member score from "assigned leader" group exercise[1,2]); peer nominations[1,2]; and cognitive ability[2]	C-UW	0.28	141	Group judgment by assessors	0.11	141
19. Work–Performance	Personal communication (1998)	Job performance		A-UW	0.35	233	Assessor judgments	0.26	112
20. Work–Performance	Personal communication (1998)	Job performance		A-UW	0.31	163	Assessor judgments	0.24	48
21. Work–Performance	Personnel communication (1998)	Job performance		A-UW	0.38	120	Assessor judgments	0.30	120

(*table continues*)

Table 1 (*continued*)

Analyses in which included	Authors (Year)	Criteria	Predictors	$Type_{mech}$	r_{mech}	N_{mech}	$Type_{clin}$	r_{clin}	N_{clin}
22. Work–Advancement	Kuncel (1999)	Job advancement	Various factors combining cognitive ability, personality, and assessment center exercises[1,2]	A-UW	0.37	270	Clinical synthesis	0.23	270
23. Work–Performance	Kuncel (1999)	Job performance	Various factors/dimensions combining constructs and methods such as cognitive ability, personality, leadership and assessment center exercises[1,2]	A-UW	0.31	270	Clinical synthesis	0.20	270
24. Work–Performance	Silzer (1984)	Job performance	Various factors/dimensions combining constructs and methods such as cognitive ability, personality, leadership and assessment center exercises[1,2]	MR	0.39	208	Clinicians' ratings based on assesse task and test files	0.37	208
25. Work–Performance	Arthur et al. (2003)	Various job-related criteria	Various factors/dimensions combining constructs and methods such as cognitive ability, personality, leadership and assessment center exercises[1,2]	MR	0.45	3,645 (131)[3]	Overall assessment center ratings	0.37	12,235 (Gaugler et al., 1987) (131)[3]

Note. $Type_{mech}$ = type of mechanical data combination; r_{mech} = observed correlation for the mechanical data combination method; N_{mech} = number of persons for whom a mechanical data combination method was used to make a prediction; $Type_{clin}$ = type of clinical data combination; r_{clin} = observed correlation for the clinical data combination method; and N_{clin} = number of persons for whom a clinical data combination method was used to make a prediction. For Analyses column, Acad. = academic; GPA = grade point average. For "Criteria" column, Ach. = achievement. For "Predictors" column, superscript 1 = used in mechanical data combination, and superscript 2 = available to clinician for clinical data combination. For "$Type_{mech}$" column, MR = multiple regression; r = correlation; DF = discriminant function; C = compositing; C-B = bootstrapped compositing; P = pooling; S = summation; A = averaging; UW = unit-weighting; DW = differential weighting; and BW = bootstrapped weighting. For "N_{mech}" and "N_{clin}" columns, superscript 3 = *N* used for meta-analysis appears in parentheses and was the median of the *N*s of the other studies in the analysis for which an *N* was known. For the studies for which the median was used as the *N* in the meta-analysis, either (a) the original source materials for the study could not be located, but we knew the effect size and other pertinent information except for the *N*, or (b) the actual *N* was so large that if it were used in the meta-analyses other than for estimating study-specific sampling error, then it would mathematically overwhelm the results.

Table 2
Meta-Analysis of Mechanical and Clinical Combinations of Predictors for Five Criteria

Criterion	N_{mech}	N_{clin}	k	r_{mech}	r_{clin}	$\sigma_{obs-mech}$	$\sigma_{obs-clin}$	$\sigma_{\rho-mech}$	$\sigma_{\rho-clin}$
Work: Job Performance	1,392	1,156	9 (5 rs, 4 Rs)	No Shrinkage (R): 0.47	0.28	0.14	0.09	0.12	0.03
				Population (ρ): 0.44	0.28	0.11	0.09		0.03
				New Sample (ρ_c): 0.40	0.28	0.08	0.09	0.05	0.03
Work: Advancement	683	683	5 (1 r, 4 Rs)	No Shrinkage (R): 0.50	0.36	0.13	0.12	0.12	0.10
				Population (ρ): 0.42	0.36	0.10	0.12		0.10
				New Sample (ρ_c): 0.36	0.36	0.11	0.12	0.08	0.10
Work: Training	188	188	2 (2 rs)	No Shrinkage (R): 0.31	0.16	0.05	0.08	0.00	0.00
				Population (ρ): 0.31	0.16	0.05	0.08	0.00	0.00
				New Sample (ρ_c): 0.31	0.16	0.05	0.08	0.00	0.00
Academic: Grade Point Average	426	471	6 (2 rs, 4 Rs)	No Shrinkage (R): 0.59	0.48	0.09	0.12	0.05	0.08
				Population (ρ): 0.58	0.48	0.09	0.12		0.08
				New Sample (ρ_c): 0.56	0.48	0.09	0.12	0.06	0.08
Academic: Non-Grade	463	161	3 (1 r, 2 Rs)	No Shrinkage (R): 0.48	0.46	0.05	0.17	0.00	0.13
				Population (ρ): 0.47	0.46	0.05	0.17		0.13
				New Sample (ρ_c): 0.46	0.46	0.05	0.17	0.03	0.13

Note. N_{mech} = number of persons for whom a mechanical data combination method was used to make a prediction; N_{clin} = number of persons for whom a clinical data combination method was used to make a prediction; k = number of samples that each contained a comparison between mechanical data combination and clinical data combination (each contained data included in the analysis); r = sample whose included effect size is a zero-order correlation; R = sample whose included effect size is a multiple correlation; No Shrinkage (R) = sample size weighted mean correlation whose multiple correlation components are observed values from the samples on which the regression equations were developed; Population (ρ) = sample size weighted mean correlation whose multiple correlation components are shrunk to the population level; New Sample (ρ_c) = sample size weighted mean correlation whose multiple correlation components are cross-validated estimates (shrunk to the level of a new sample from the same population); r_{mech} = sample size weighted mean correlation for the mechanical data combination methods (composite of multiple and/or zero-order correlations); r_{clin} = sample size weighted mean observed correlation for the clinical data combination methods; $\sigma_{obs-mech}$ = sample size weighted observed standard deviation of the correlations for mechanical data combination; $\sigma_{obs-clin}$ = sample size weighted observed standard deviation of the correlations for clinical data combination; $\sigma_{\rho-mech}$ = standard deviation of correlations for mechanical data combination after removing sampling error variance; $\sigma_{\rho-clin}$ = standard deviation of correlations for clinical data combination after removing sampling error variance.

made by organizational members (e.g., what predictors to use, what predictors to ignore, to whom to extend a job offer) and job applicants (e.g., the decision to apply, what level of effort to exert during selection, the decision to accept a job offer). The present study highlights the importance of this focus because the predictive power of the selection/admissions systems is affected by human judgment, and is not the same as a weighted sum of their parts. Yet, the importance of the judgment and decision making framework extends beyond the present study and upends traditional validation research in some critical and radical ways. We discuss three implications of the Lens Model for selection that are important for understanding the limitations of the present study and key directions for future research. Note that this is far from an exhaustive list.

The Lens Model can be extended to include the decision to hire, decisions to apply/accept, and the effect of hiring on subsequent performance. Within a decision-to-hire framework, correlations between predictors and observed performance in a validation study *do not* necessarily reflect the utility of a predictor when used in a hiring decision (even in the simplified case where all job offers are accepted). The judgment to extend an offer can have no relationship with predictor scores even though the predictor cues are associated with subsequent job performance. That is, a predictor can be discounted when hiring employees and have no effect on hiring decisions. Within this framework, a traditionally valid ($r > 0$) predictor that does not affect hiring judgments has negative utility due to the cost of using the predictor. Obtaining a non-zero correlation between a predictor and subsequent job-performance does not tell us if it favorably influences hiring decisions. The correlation is only an unambiguous measure of predictive validity if the predictor is used in a strict top down selection format. If hiring judgments deviate from the top down selection decisions, then predictive power and utility will differ.

Second, incremental predictive power as measured by multiple regression analyses will typically reflect a rarely occurring (and often idealized) setting where decision making is based on strict differential weighting and top down selection. In contrast, within a judgment framework, redundant predictors ($\Delta R = 0$) can improve prediction by pushing out or reducing a human judge's emphasis on invalid cues. Double counting a redundant predictor helps if it makes one ignore invalid variance in forming a judgment. The model suggests that face validity for the decision maker (often considered a side issue or external marketing concern in selection research) becomes a critical feature as it likely influences use and subjective weighting of decision aids. Put simply, no matter how valid a predictor is, if it is not liked by decision makers, it likely will not improve the decision quality. The same issue applies to experts combining information from many cues.

Third, a potential applicant's decision to apply can dramatically affect the nature of the pool and, therefore, the expected average performance of new workers after making the hiring decision. For example, Kuncel and Klieger (2007) reported that when applicants had information about the likelihood of acceptance they generally chose to avoid applying to law schools for which they were either under or over qualified. The resulting applicant pools across law

schools differed dramatically as a result. As selection system information becomes public, applicant pools may shift depending on their perception of the system.

Practice Suggestions

While recognizing that a strong preference for expert judgment makes a complete change in practice unlikely, a number of methods could be adopted that could yield immediate benefits.[5] First, in cases with many applicants, mechanical methods could be used to screen all but a final pool. Second, experts could use mechanically combined data as an anchor and make limited (or consensus based) adjustments. Third, documenting the reason for deviations from mechanically combined scores makes the decision public and would permit follow up research and feedback. Fourth, both expert combined and mechanically combined scores could be presented to decision-makers. This fourth approach also allows for a narrative explaining the difference. Finally, given the previous literature, the most likely source of the difference is lower reliability for the clinical approaches (i.e., less consistent and more fraught with unsystematic errors). Therefore, the method with the most potential for improved predictive power would be to average across multiple raters even if secondary (and possibly less involved) raters were given a lower weight in the final assessment.

Research on each of these suggestions would be invaluable particularly if embedded in the broader judgment and decision making framework outlined in the introduction and discussion. It is possible that less valid data combination methods (in a correlational sense) have a larger positive effect on end user decision making due to greater face validity and acceptability. We believe research on three general questions are crucial. First, why does expert judgment result in lower correlations? Second, why do decision makers use or ignore information in decision making? Third, what alternative methods improve predictive power while retaining acceptability? Finally, it is not unreasonable to believe that experts have important insights. Unfortunately, it appears that this comes at too high a cost. Therefore, what can be done to capture insights while avoiding validity damaging inconsistency?

For rare and highly complex jobs, future research should consider adopting a forecasting framework where experts make specific and verifiable predictions about the future behavior of assessees. This framework will allow for the accumulation of data in small *N* settings and advance the field.

Highhouse (2008a) noted "arguments in favor of holistic assessment, nevertheless, sometimes take on a faith-based quality and fail to acknowledge the preponderance of the evidence" (p. 375). Consistent with the preponderance of the evidence, this meta-analysis found and quantified that a consistent and substantial loss of information occurs when data are combined clinically—even by experts who are knowledgeable about the jobs and organizations in question—across multiple criteria and work or academic settings.

On the positive side, it is clear that psychological assessments do predict subsequent performance across outcomes and domains. We do useful work. Also clear is that improvements can be made. The results do not mean that experts are unimportant. Again, the literature demonstrates that data combination is best done mechanically while information collection can be done quite effectively by experts. Overall, the time of experts would be best invested in collecting job relevant information about candidates or working on subsequent development rather than judgment based data combination.

Although the widespread replacement of clinical methods with mechanical methods is unlikely in the foreseeable future, we see this study's findings as a call to find hybrid methods of data combination that improve on expert judgment while remaining acceptable to end users. We take a pragmatic view of this issue. Surveys have suggested that although 2% of people involved in individual assessment make use of purely mechanical methods, close to half report using methods that combine statistical and holistic methods (Ryan & Sackett, 1987). Although the nature and effectiveness of these, likely varied, approaches is unknown, it appears that there is room to develop methods that move the mean upward while retaining approaches that are attractive to professionals and end users. Evidence-based practice can benefit from keeping the results of this meta-analysis in mind when developing and utilizing selection and admission systems.

[5] With the possible exception of organizations concerned with equal employment opportunity (EEO) compliance, which will use fixed weights to avoid charges of disparate treatment.

References

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125–153. doi:10.1111/j.1744-6570.2003.tb00146.x

Borman, W. C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology, 67,* 3–9.

Borman, W. C., & Brush, D. H. (1993). More progress toward a taxonomy of managerial performance requirement. *Human Performance, 6,* 1–21. doi:10.1207/s15327043hup0601_1

Browne, M. W. (1975). Predictive validity of a linear regression equation. *British Journal of Mathematical and Statistical Psychology, 28,* 79–87. doi:10.1111/j.2044-8317.1975.tb00550.x

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193–217.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. (1970). *Managerial behavior, performance, and effectiveness*. New York, NY: McGraw-Hill.

Cattin, P. (1980). Estimation of predictive power of a regression model. *Journal of Applied Psychology, 65,* 407–414. doi:10.1037/0021-9010.65.4.407

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement, 2,* 595–607. doi:10.1177/014662167800200414

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26,* 180–188. doi:10.1037/h0030868

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582. doi:10.1037/0003-066X.34.7.571

Feltham, R. (1988). Assessment centre decision making: Judgmental vs. mechanical. *Journal of Occupational Psychology, 61,* 237–241. doi:10.1111/j.2044-8325.1988.tb00287.x

Flanagan, J. C. (1951). Defining the requirements of the executive's job. *Personnel, 28,* 28–35.

Fowler, R. L. (1986). Confidence intervals for the cross-validated multiple correlation in predictive regression models. *Journal of Applied Psychology, 71,* 318–322. doi:10.1037/0021-9010.71.2.318

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511. doi:10.1037/0021-9010.72.3.493

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73,* 422–432.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323. doi:10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30. doi:10.1037/1040-3590.12.1.19

Hammond, K. R., & Summers, D. A. (1972). Cognitive control. *Psychological Review, 79,* 58–67.

Highhouse, S. (2008a). Facts are stubborn things. *Industrial and Organizational Psychology, 1,* 373–376. doi:10.1111/j.1754-9434.2008.00069.x

Highhouse, S. (2008b). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology, 1,* 333–342. doi:10.1111/j.1754-9434.2008.00058.x

Huck, J. R. (1974). *Determinants of assessment center ratings for White and Black females and the relationship of the dimensions to subsequent performance effectiveness* (Unpublished doctoral dissertation). Wayne State University, Detroit, MI.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

Jeanneret, R., & Silzer, R. (1998). An overview of individual psychological assessment. In R. Jeanneret & R. Silzer (Eds.), *Individual psychological assessment* (pp. 3–26). San Francisco, CA: Jossey-Bass.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment A meta-analysis of lens model studies. *Psychological Bulletin, 134,* 404–426. doi:10.1037/0033-2909.134.3.404

Kuncel, N. R. (1999). *Maximizing validity and utility with multiple predictors* (Unpublished master's thesis). University of Minnesota, Minneapolis.

Kuncel, N. R. (2008). Some new (and old) suggestions for improving personnel selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 343–346. doi:10.1111/j.1754-9434.2008.00059.x

Kuncel, N. R., & Klieger, D. M. (2007). Applicant patterns when applicants know the odds: Implications for selection research and practice. *Journal of Applied Psychology, 92,* 586–593. doi:10.1037/0021-9010.92.2.586

Lewis, E. C., & MacKinney, A. C. (1961). Counselor vs. statistical predictions of job satisfaction in engineering. *Journal of Counseling Psychology, 8,* 224–230. doi:10.1037/h0043207

Mendoza, J. L., & Stafford, K. L. (2001). Confidence intervals, power calculation, and sample size estimation for the squared multiple correlation coefficient under the fixed and random regression models: A computer program and useful standard tables. *Educational and Psychological Measurement, 61,* 650–667. doi:10.1177/00131640121971419

Mitchel, J. O. (1975). Assessment center validity: A longitudinal study. *Journal of Applied Psychology, 60,* 573–579. doi:10.1037/0021-9010.60.5.573

Neidich, A. (1968). *Honors Selection Study 1966–67*. Columbia: University of South Carolina.

Nunnally, J. C. (1978). *Psychomteric theory*. New York, NY: McGraw-Hill.

Nystedt, L., & Magnusson, D. (1972). Predictive efficiency as a function of amount of information. *Multivariate Behavioral Research, 7,* 441–450. doi:10.1207/s15327906mbr0704_2

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29,* 201–211. doi:10.1214/aoms/1177706717

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement, 23,* 99–115. doi:10.1177/01466219922031220

Robertson, M., & Hall, E. (1964). Predicting success in graduate school. *Journal of General Psychology, 71,* 359–365.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlation: A clarification. *Psychological Bulletin, 85,* 1348–1351. doi:10.1037/0033-2909.85.6.1348

Ryan, A. M., & Sackett, P. R. (1987). A survey of individual assessment practices by I/O psychologists. *Personnel Psychology, 40,* 455–488. doi:10.1111/j.1744-6570.1987.tb00610.x

Sarbin, T. R. (1943). A contribution to the study of actuarial and individual methods of predictions. *American Journal of Sociology, 48,* 593–602. doi:10.1086/219248

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66,* 178–200. doi:10.1037/h0023624

Shieh, G. (2008). Improved shrinkage estimation of squared multiple correlation coefficient and squared cross-validity coefficient. *Organizational Research Methods, 11,* 387–407. doi:10.1177/1094428106292901

Silzer, R. F. (1984). *Clinical and statistical prediction in a management assessment center*. Minneapolis: University of Minnesota.

Stuit, D. B. (Ed.). (1947). *Personnel research and test development in the Bureau of Naval Personnel*. Princeton, NJ: Princeton University Press.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical validity of tests in selection. *Journal of Applied Psychology, 23,* 565–578. doi:10.1037/h0057079

Truesdell, A. B., & Bath, J. A. (1957). Clinical and actuarial predictions of academic survival and attrition. *Journal of Counseling Psychology, 4,* 50–53. doi:10.1037/h0044494

Tziner, A., & Dolan, S. (1982). Validity of an assessment center for identifying future female officers in the military. *Journal of Applied Psychology, 67,* 728–736. doi:10.1037/0021-9010.67.6.728

Watley, D. J., & Vance, F. L. (1964). *Clinical versus actuarial prediction of college achievement and leadership activity* (Book No. Cooperative Research Project No. 2202). Minneapolis: University of Minnesota.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics, 2,* 440–457. doi:10.1214/aoms/1177732951

Wiggins, N., & Kohen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology, 19,* 100–106. doi:10.1037/h0031147

Wollowick, H. B., & McNamara, W. J. (1969). Relationship of the components of an assessment center to management success. *Journal of Applied Psychology, 53,* 348–352. doi:10.1037/h0028102

Yin, P., & Fan, X. (2001). Estimating R^2 shrinkage in multiple regression: A comparison of different analytical methods. *Journal of Experimental Education, 69,* 203–224. doi:10.1080/00220970109600656

(*Appendix follows*)

Appendix

Cross-Validation Estimation Method

To solve for the point and sampling error estimates, one usually has to solve first for ρ and ω, because among the values in the equations they usually are the only unknowns (see Table A1). Calculating ρ (or at least an estimate of it) necessitates determining whether the predictor model in question is fixed or random (see Cattin, 1980). In fixed predictor models (FPMs), the predictors in the equation are the only predictors that could have been used to address the research question. In random predictor models (RPMs), the predictors in the equation are just a sample of the predictors that could have been used to address the research question. For FPMs, one should use Wherry's (1931) formula, although it may be slightly biased (Cattin, 1980). For RPMs, one should use a version of Olkin and Pratt's (1958) formula (Cattin, 1980; Shieh, 2008; Yin & Fang, 2001). Most questions in social science use the RPM rather than the FPM (Cattin, 1980), and the RPM seemed more appropriate for the studies being meta-analyzed. Shieh (2008) found that a slightly modified version of Olkin and Pratt's (1958) formula for estimating ρ performed best in simulations (see Table A1, Formula 2). To solve for ω, one uses Browne's (1975) Equation 2.8 recommended by Cattin (1980) and Shieh (2008).

Table A1
Key Equations for Cross-Validation Estimates

Formula	Source(s)
1. $\varepsilon(\omega^2) = \omega^2 - \frac{2(N-p-2)(N-2p-6)(p-1)\rho^4(1-\rho^2)^2}{(N-p-4)\{(N-2p-2)\rho^2+\rho\}^3} + o(\{N-p\}^{-1})$	Browne's (1975) Equation 2.10
2. $\hat{\rho}_P^2(R^2) = 1 - \frac{N-3}{N-p-1}(1-R^2)\left\{1 + \frac{2(1-R^2)}{N-p-2.3}\right\}$ where $\hat{\rho}_P^2(R^2) = 0$ if $\hat{\rho}_P^2(R^2) < 0$	Shieh (2008) (based on Olkin & Pratt, 1958)
3. $\omega^2 = \frac{(N-p-3)\rho^4 + \rho^2}{(N-2p-2)\rho^2 + p}$	Browne's (1975) Equation 2.8; Cattin (1980); Shieh (2008)
4. $\text{var}(\omega^2) = \frac{2(N-p-2)(p-1)\rho^4(1-\rho^2)^2\{2(N-p-5)\rho^2+1-(N-2p-6)\omega^2\}}{(N-p-4)\{(N-2p-2)\rho^2+p\}^3} + o(\{N-p\}^{-1})$	Browne's (1975) Equation 2.11

Note. $\omega = \rho_c$ = population cross-validated multiple correlation; $\varepsilon(\omega^2) = \hat{\rho}_c^2$ = estimated cross-validated multiple correlation, squared; $var(\omega^2) = var(\hat{\rho}_c{}^2)$ = variance of estimated cross-validated squared multiple correlation; ρ = population multiple correlation; R^2 = observed multiple correlation, squared (a.k.a., observed coefficient of determination); $\hat{\rho}_P^2(R^2) = \hat{\rho}^2$ = estimated population multiple correlation, squared; N = number of observations; p = number of predictor variables; o = "little o" = a function describing the limit on error = how far off one's obtained value can be.

Received January 22, 2009
Revision received February 4, 2013
Accepted March 6, 2013 ■