Personnel Assessment and Decisions

Volume 6
Issue 2 *Applications of Judgment and Decision Making to Problems in Personnel Assessment*

Article 2

2020

Pushing the Limits for Judgmental Consistency: Comparing Random Weighting Schemes with Expert Judgments

Martin C. Yu
HumRRO

Nathan R. Kuncel
University of Minnesota - Twin Cities

Recommended Citation

Yu, Martin C. and Kuncel, Nathan R. (2020) "Pushing the Limits for Judgmental Consistency: Comparing Random Weighting Schemes with Expert Judgments," *Personnel Assessment and Decisions*: Number 6 : Iss. 2 , Article 2.
DOI: https://doi.org/10.25035/pad.2020.02.002
Available at: https://scholarworks.bgsu.edu/pad/vol6/iss2/2

Pushing the Limits for Judgmental Consistency: Comparing Random Weighting Schemes With Expert Judgments

Martin C. Yu[1] and Nathan R. Kuncel[2]

1. HumRRO
2. University of Minnesota - Twin Cities

KEYWORDS

judgment and decision making, personnel selection

ABSTRACT

Consistent use of information has been identified as a critical issue that can undermine expert predictions. Using three personnel assessment datasets, we conduct Monte Carlo simulations to compare the accuracy of expert judgements for predicting the job performance of managers against four different weighting schemes: consistent random weights, completely random weights, unit weights, and optimal weights. Expert accuracy fell within the completely random weight distribution in two samples and at the low end of the consistent random weight distribution in one sample. In other words, consistent random weights reliably outperformed expert judgment for hiring decisions across three datasets with a total sample size of 847. We see this as a call to develop decision making systems that help control consistency or to manage consistency by aggregating multiple expert judgments.

Making correct hiring decisions can have far-reaching implications for ensuring organizational productivity and success. Evaluating job candidates often boils down to combining various pieces of information (e.g., simulations, roleplaying, interviews, and intelligence and personality tests) on job-related dimensions such as communication and leadership skills to form a judgment about candidates' potential fit and performance on the job. Therefore, understanding the psychological and cognitive processes behind how these judgments are made would be imperative to maximize the predictive validity of our selection systems.

When combining information to make a judgment, there are typically two general ways to go about doing so. The first is the use of mechanical methods where information is combined statistically/algorithmically using some predefined mathematical formula. The second is the use of clinical judgment, sometimes referred to as human, expert, holistic, intuitive, or subjective judgment. Here, expert judgments are made using the personal intuition or insight of the person making the judgment. In comparing the efficacy of these two approaches, one of the most consistent findings in the judgment and decision-making literature has been that mechanical, algorithmic methods tend to outperform those made using expert judgment. This has been the case across a variety of decision making scenarios and across a variety of mechanical methods, including those as simple as unit weighting where predictor scores are simply added up (Grove & Meehl, 1996; Grove et al., 2000), and this holds true in the context of hiring and admissions decisions (Kuncel et al., 2013). Although certain predictors can be highly valid, ultimately the method used to combine predictor information can serve to either maximize or limit the accuracy of the prediction system. As Hastie and Dawes (2001) stated:

> a substantial amount of time and other resources is squandered on expert judgments that could be made more equitably, more efficiently, and more accurately by the statistical models we humans construct than by we humans alone. (p. 63)

On average, mechanical methods outperform human judgment, so ideally, predictive power would be maximized by using mechanical methods as much as possible. However, there are considerations that may limit the use of algorithmic judgment in reality. Although the use of algorithmic

Corresponding author:
Martin C. Yu
66 Canal Center Plaza, Suite 700
Alexandria, VA 22314
Email: myu@humrro.org

and artificial intelligence systems in personnel selection is expected to increase (Putka & Dorsey, 2019), in practice, decisions made using clinical judgments have long been preferred over those made mechanically (e.g., Highhouse, 2008; Jeanneret & Silzer, 1998; Ryan & Sackett, 1987; Slaughter & Kausel, 2013). In addition to this preference, there are reasonable concerns regarding purely relying on mechanical methods, such as face validity, adoption of decision aids, and candidate reactions (e.g., Diab et al., 2011; Eastwood et al., 2012; Kuncel, 2018). Simply put, people tend not to like being reduced to a set of numbers in and a number out, and decision makers often feel uncomfortable solely relying on an algorithm. Given these considerations, it may be more practical to improve and support human judgment rather than to try to replace it with mechanical methods. To that end, it will be necessary to better understand why mechanical methods often have an advantage over clinical judgment in hiring decision making.

The lens model (Brunswik, 1952; Hammond, 1955) has been a well-established framework for analyzing human judgmental processes, and research with this approach has demonstrated that expert judgment is outperformed by mechanical methods of data combination and that this is explained by both use of inaccurate weights in combining predictor information (low cue sensitivity) as well as inconsistent use of these weighting policies (low cognitive control; Karelaia & Hogarth, 2008). The lens model equation (Tucker, 1964) describes the components that influences the accuracy of human judgment:

$$r_a = GR_e\,R_s + C\sqrt{(1 - R_e^2)(1 - R_s^2)}$$

Here, human judgment accuracy (r_a) is a function of the accuracy of the judge's policy for combining predictor information (i.e., cue sensitivity; G), the predictability of the criterion (i.e., environmental predictability; R_e), the consistency with which the judge applies his or her own judgmental policy (i.e., cognitive control; R_s), and any random or systematic error not captured by the lens model (i.e., unmodeled knowledge; C). When judges use optimal cue weights that reflect the actual predictor–criterion relationships and/or use their weighting policy consistently, their judgmental accuracy will increase accordingly. The lens model has been used to examine multiple questions including topics in the world of work (e.g., Schmitt et al., 1986).

Additionally, the human judge will best the mechanical model in prediction to the extent that there is unmodeled knowledge for which the judge is able to consistently and accurately incorporate into his or her judgments. For example, if expert judges are able to validly account for red flags, interactions, or other non-linearities that may not be captured by a mechanical model, they would be able to maintain their ability to make valid predictions even though their judgmental policies will likely vary from case to case as they incorporate different pieces of information into their judgments or weigh information cues differently. To the extent that the human judge is able to validly account for any information not accounted for by a mechanical model, this would constitute valid use of expert insight to improve prediction.

In the context of employment hiring decisions, there has been a heavy reliance on expert judgment in individual assessments due to what is essentially a belief that there is a substantial amount of unmodeled knowledge on which an expert judge is able to capitalize. Silzer and Jeanneret (2011) produced what is, to date, probably the most extensive description of all the skills and abilities that expert assessors supposedly bring to the table when conducting individual assessments. They make a number of bold claims about the use of expert judgment, including that expert assessors:

- "are accurate observers of behavior … can see and hear behavior in their observations of an individual that can provide useful and sometimes critical pieces of information to rating the individual on key dimensions" (p. 276)
- "can also formulate and test hypotheses about the individual. Using an analytical approach, they can probe and collect additional information relevant to a concern or a dimension" (p. 276)
- "can understand specific behavioral data points while also seeing larger behavioral patterns and psychological constructs" (p. 276)
- "can complete both normative and ipsative interpretations for the same variables for the same assessee that leads to a fuller understanding of that individual … a process that would be virtually impossible to complete in some mechanical or statistical manner" (p. 276)
- "can accurately sort behavior into key performance-related dimensions" (p. 277)
- "can integrate information and accurately rate an individual on specific performance dimensions" (p. 277)
- "can consider a range of behavior and determine how relevant the behavior is to later performance effectiveness" (p. 277).

In summary, Silzer and Jeanneret (2011) claim that expert assessors are able to effectively exercise their intuitive judgment to validly integrate information in complex ways. However, their assertions have been challenged (Kuncel & Highhouse, 2011) and are not well supported by empirical evidence. Lens model research has shown that unmodeled knowledge tends to be very low, leaving little room, if any, for the expert to improve over a mechanical model (Karelaia & Hogarth, 2008).

Given that there is little unmodeled knowledge to exploit, the main drivers of judgmental accuracy as indicated by the lens model would then be the use of accurate predictor weights and the consistency with which these weights are applied across judgmental cases. The issue with human judgment here is that even experts can lack insight into their own judgmental policies (Hastie & Dawes, 2001). Clearly, it would be difficult to apply predictor weights accurately and consistently without a firm grasp of one's own judgmental policy. This is where mechanical methods of judgment shine because they can be programmed to consistently apply a single set of decision rules (e.g., predictor weights) across every single judgmental case. With mechanical methods, we know specifically what judgmental policy is being used and that it is being applied consistently.

This raises the question of whether it is the use of accurate (optimal) weighting schemes or the consistency with which a weighting scheme is applied that drives the predictive power of a judgmental method, or if they are equally influential. Past evidence suggests that consistency is more important than optimality. Linear models are robust (Dawes, 1979), meaning that changes in predictor weights do not drastically impact their predictive power as long as the signs on the weights do not change (i.e., positive weights stay positive, and negative weights stay negative). In multiple regression with three or more predictors, an infinite class of alternate regression weights (i.e., fungible weights) can be generated that yield a predictive validity approaching that of the optimal set of predictor weights (Waller, 2008). Moreover, Dawes and Corrigan (1974) found that, on average, a mechanical combination using random positive weights applied consistently across all judgmental cases was able to match or outperform human judges across five different judgment and decision-making scenarios.

This study uses three real assessment data sets with job performance measures to examine how different simulated weighting schemes compares to expert judgment. This is an extensive extension of Dawes and Corrigan (1974) to more thoroughly study the degree to which inconsistency in combining information when making multiple judgments is detrimental to the predictive validity of expert judgment. Because the judgmental processes involve two aspects of data combination – the optimality of the data combination policy and the consistency with which the policy is applied – it would be necessary to tease apart consistency from optimality if the effects of consistency are to be studied. This can be done by examining random weighting schemes as there is no expectation of optimality and by pitting expert judgment against random weights in combining predictor information. When the intent is to make the most accurate judgment possible, randomly weighting information cues to make a judgment is the complete opposite of using a set of optimal regression weights.

There are two forms of random weighting that warrant consideration. The first form is the one used by Dawes and Corrigan (1974), where a set of random weights is generated and applied consistently to every single judgmental case. In a simulation study, this is repeated many times so that the average validity of consistent use of random weights can be estimated. The second form is inconsistent weighting, where a set of random weights is generated for every single judgmental case. Here, no two judgments are combined using the same weighting policy (unless by coincidence). Again, this process is repeated many times to estimate the average validity of random weighting. With consistent random weights, there is no expectation of optimality, but there is an expectation of consistency. With inconsistent weights on the other hand, there is no expectation of either optimality or consistency.

Beyond Dawes and Corrigan (1974), where only the average validity of consistent random weights was evaluated, the distribution of the predictive validities of random methods across all simulated iterations should also be examined and used to benchmark the predictive validities of non-random methods (e.g., simple unit weights and optimal weights) against expert judgment. Answering these questions will provide a stronger theoretical understanding regarding why mechanical methods tend to perform better than subjective expert judgment and will also provide practical insight into possible means of supporting and improving expert judgment.

In this study, we build on previous research and provide a more detailed examination of random weighting schemes and their implications for understanding the importance of exercising consistency in judgment and decision making processes. Using data from personnel assessments conducted at two separate companies by an international management consulting firm, we run Monte Carlo simulations for the application of random weights consistently across all judgmental cases and for the application of random weights inconsistently across all judgmental cases. The validity of these composite scores made using these random methods for predicting assessment candidates' future job performance are compared to the validity of judgments made using non-random methods of data combination: subjective expert judgment, unit weighting via simple sums, and optimal weighting. We examine not only the average validity of these random methods but also their variability across many simulated iterations. Based on these distributions, we would then be able to determine the extent to which non-random methods of prediction outperform or do not outperform these random methods.

Using this analytical approach, optimal regression weighting is expected to outperform the random weighting methods in almost all cases, save any case where the random weights coincidentally approach the optimal weights. Unit weighting via simple sums is also expected to outperform inconsistent weighting in a large majority of cases. If sampling error in generating the consistent random weighting schemes is distributed evenly about the unit weights, unit weighting would likely be better than consistent random weighting about half the time and worse the other half. Consistent application of a single set of random weights across all judgmental cases should yield more valid predictions of job performance compared to inconsistent weighting.

Given the importance of consistency as discussed previously, consistent random weighting is expected to outperform expert judgment in an overwhelming majority of cases. If we also see that inconsistent weights mirror the predictive power of expert judges, then we have strong evidence that the judges are using information very inconsistently and that this inconsistency in combining information does not reflect utilizing expert insight and strategies specific to individuals, contexts, or jobs that improve their judgments.

METHOD

Sample

Three archival assessment validation datasets were obtained from an international management consulting firm[1]: (a) Company A, a financial services provider (231 candidates evaluated by 26 assessors between 1994 and 1997); (b) Company B, a food retailer, Sample 1 (195 candidates evaluated by 23 assessors between 1980 and 1988); and (c) Company B, Sample 2 (421 candidates evaluated by 30 assessors between 1989 and 1999). Sample 1 and Sample 2 from Company B were obtained from separate validation studies. Candidates were evaluated for management positions by doctoral-level psychologists trained in conducting managerial hiring assessments.

Based on their performance on a mix of in-basket, interviews, leaderless group discussions, personality test, and cognitive ability test, candidates were rated on seven assessment dimensions: adjustment, administration, communication, interpersonal, judgment, leadership, and motivation. Using these dimension ratings, the assessors then combined each of their candidates' ratings on these dimensions into an overall assessment rating based on person–job fit, such that employees who better fit with the job are expected to perform better on the job (e.g., Kristof-Brown et al., 2005). Supervisory ratings of job performance are used as the criterion variable.

Missing data were handled by multiple imputation with predictive mean matching (Schenker & Taylor, 1996). This method randomly samples donor values from neighboring observations that has a predicted value closest to the predicted value of the missing value. As it samples values from existing data, it maintains the plausibility of the imputed values compared to other regression-based methods. Using the MICE (Multiple Imputation by Chained Equations) package in R (van Buuren & Groothuis-Oudshoorn, 2011), five imputed datasets were generated for each of the three archival datasets, and analyses for each archival dataset were pooled across all five imputed datasets.

Analyses were conducted using both listwise deletion and multiple imputation. Conclusions were the same for both methods of handling missing data. Because listwise deletion is the less preferable option (Newman, 2014) and for the sake of brevity, only results obtained via multiple imputation will be presented.

Analyses

The analyses described in this section were conducted separately using each of the three validation datasets. To simulate the use of random weights applied consistently, a set of seven weights were randomly sampled from a uniform distribution that ranged from 0 to 0.5, inclusive. The weights sampled were constrained to be positive as the pairwise relationships between each dimension rating and the overall assessment rating were expected to be positive, and maintaining the same sign (i.e., positive or negative) for the weight is important for maintaining the predictive relationships (Dawes, 1979). They were also constrained to the 0 to 0.5 range to simulate correlational weights between individual assessments and job performance (Morris et al., 2015), and to limit the degree to which each dimension could be differentially weighted. This same set of random weights was then used to linearly combine each candidate's seven assessment dimension ratings into an overall assessment rating. These overall ratings were then correlated with the candidates' supervisory ratings of job performance as a measure of the predictive validity of applying a set of random weights consistently. This process was iterated 10,000 times, generating a total of 10,000 correlations as validity coefficients. Table 1 presents an example of a consistent random weighting scheme.

To simulate the use of inconsistent weights for each candidate, the dimension ratings for each candidate are linearly combined into an overall rating using a set of seven weights that were randomly sampled from a uniform distribution that ranges from 0 to 0.5, inclusive. As described previously, the weights were constrained to be positive. A new set of seven random weights was generated to combine the dimension ratings of each candidate into overall assessment ratings. In this case, no two candidates were evaluated using the exact same weighting scheme (unless by coincidence). Again, these overall ratings were then correlated with the candidates' supervisory ratings of job performance as a measure of the predictive validity of applying inconsistent weights. This process is iterated 10,000 times, generating a total of 10,000 correlations as validity coefficients. An

[1] The identity of the consulting firm is kept anonymous as the results of this study do not paint a positive picture. Their willingness to share the data that make this study possible is much appreciated. For readers who are somehow able to guess the identity of this firm, please note that these data are considered legacy data that do not necessarily reflect their current assessment practices.

example of an inconsistent weighting scheme is presented in Table 1.

To provide points of comparison with non-random methods, the predictive validities of overall ratings made using non-random methods – expert judgment, simple sums, and optimal weighting – were computed. First, the overall assessment ratings made using the assessors' expert judgment were correlated with the supervisory ratings of job performance as a measure of the predictive validity of expert judgment. Second, simple sum overall ratings were calculated by adding up the dimension ratings. Correlating this with the candidates' job performance ratings yielded the predictive validity of a unit weighted via simple sums composite. Last, optimally weighted overall ratings were calculated by first obtaining the optimal weights by extracting the regression coefficients from an ordinary least squares multiple linear regression model using the candidates' dimension ratings to predict their job performance. Each candidates' dimension ratings were then linearly combined using these optimal weights into an optimally weighted composite. Correlating this composite score with their job performance yielded the predictive validity of an optimally weighted composite.

RESULTS

Figure 1 displays results for analyses using the Company A data. Figure 2 displays results for analyses using the Company B, Sample 1 data; and Figure 3 displays results for analyses using the Company B, Sample 2 data.

Comparing the non-random methods in predicting supervisory ratings of job performance at Company A, overall ratings made using optimal weights (*r* = .25) were better predictors than those made using simple sums (*r* = .19), which in turn performed about the same as those made using clinical expert judgment (*r* = .17). In Company B, Sample 1, optimal weights (*r* = .40) were better than unit weights (*r* = .33), which were better than expert judgment (*r* = .16), and a similar pattern was found in Company B, Sample 2 where optimal weights (*r* = .30) were better than unit weights (*r* =.22), which were better than expert judgment (*r* = .13).

When the overall ratings computed using random methods were used to predict job performance at Company A, across 10,000 iterations, random weights applied consistently across candidates had a mean predictive validity of *r* = .18 (SD = .02) and ranged from *r* = .10 to .22. Random weights applied consistently outperformed expert judgments in 76.83% of the iterations, simple sums in 39.40% of the iterations, and never outperformed optimal weights. Inconsistent weighting across candidates had a mean validity of *r* = .09 (*SD* = .02), and ranged from *r* = -.01 to .19. Inconsistent weights never outperformed expert judgment, simple sums, or optimal weights. 69.85% of the iterations for inconsistent weights were outperformed by all of the iterations for random weights applied consistently.

At Company B, Sample 1 across 10,000 iterations, random weights applied consistently across candidates had a mean validity of *r* = .34 (*SD* = .03) and ranged from *r* = .20 to .40. Random weights applied consistently outperformed expert judgments in 100% of the iterations, simple sums in 32.96% of the iterations, and never outperformed optimal weights. Inconsistent weighting across candidates had a mean validity of *r* = .16 (*SD* = .03) and ranged from *r* = .05 to .27. Inconsistent weights outperformed expert judgments in 8.49% of the iterations but never outperformed simple sums or optimal weights. 94.05% of the iterations for inconsistent weights were outperformed by all of the iterations for random weights applied consistently.

At Company B, Sample 2 across 10,000 iterations, random weights applied consistently across candidates had a mean validity of *r* = .24 (*SD* = .02) and ranged from *r* = .15 to .29. Random weights applied consistently outperformed expert judgments in 100% of the iterations, simple sums in 36.12% of the iterations, and never outperformed optimal weights. Inconsistent weighting across candidates had a mean validity of *r* = .12 (*SD* = .02) and ranged from *r* = .05 to .20. Inconsistent weights outperformed expert judgments

TABLE 1.
Example Consistent and Inconsistent Random Weighting Schemes

	Consistent weights				Inconsistent weights			
Candidate	W1	W2	…	W7	W1	W2	…	W7
1	.02	.36		.19	.35	.45		.12
2	.02	.36		.19	.15	.11		.28
3	.02	.36	…	.19	.43	.06	…	.33
4	.02	.36		.19	.04	.41		.31
5	.02	.36		.19	.22	.17		.09

Note. Randomly generated weights were constrained to be positive values between 0 and 0.5 as pairwise predictor–criterion relationships were expected to be positive.

FIGURE 1.

Density distributions of validities (10,000 iterations each) at Company A of predictor scores combined using random positive weights applied consistently (top plot) or inconsistently (bottom plot) across all candidates. Vertical lines are validities at Company A of non-random methods of data combination: expert judgment (solid line), unit weighting via simple sums (dashed line), and optimal weighting (dotted line).

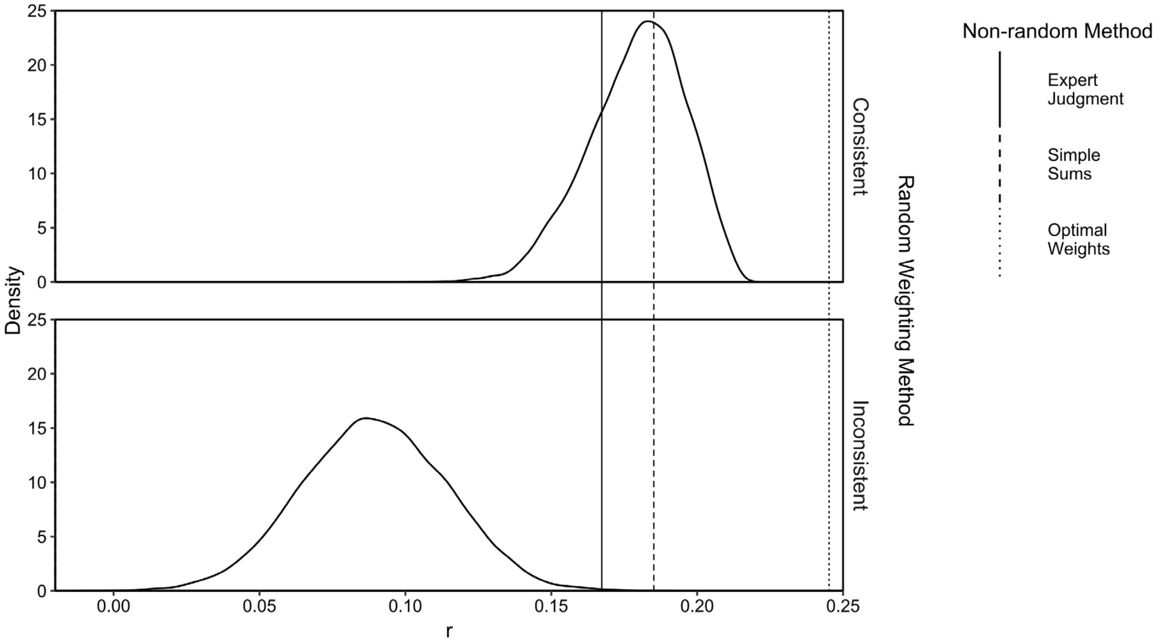


FIGURE 2.

Density distributions of validities (10,000 iterations each) at Company B, Sample 1 of predictor scores combined using random positive weights applied consistently (top plot) or inconsistently (bottom plot) across all candidates. Vertical lines are validities at Company A of non-random methods of data combination: expert judgment (solid line), unit weighting via simple sums (dashed line), and optimal weighting (dotted line).

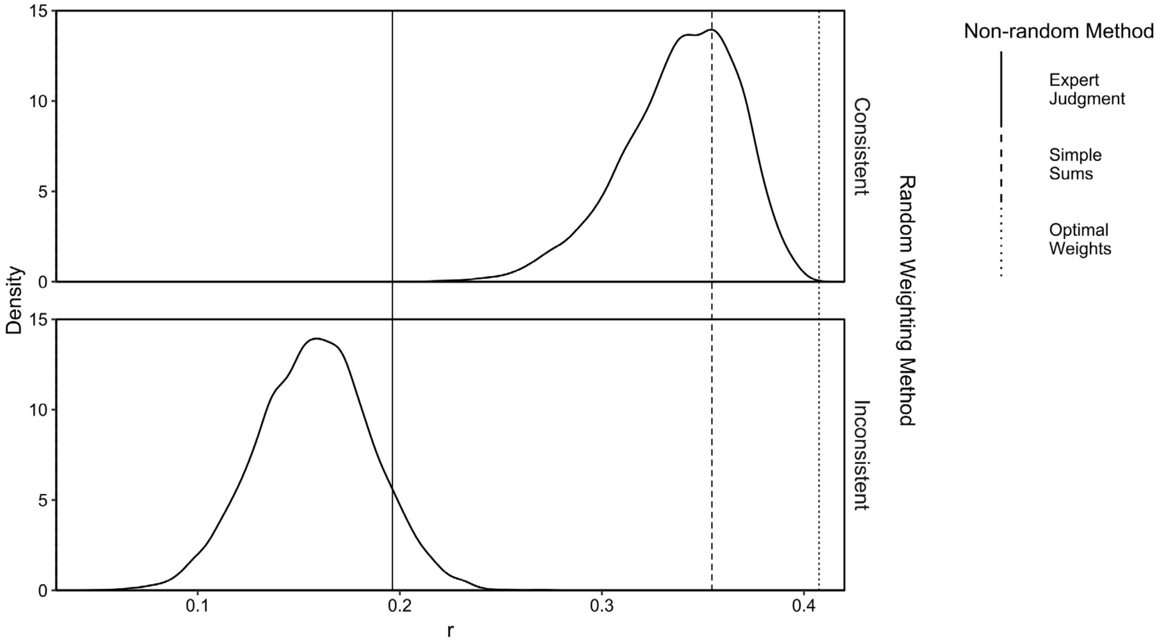
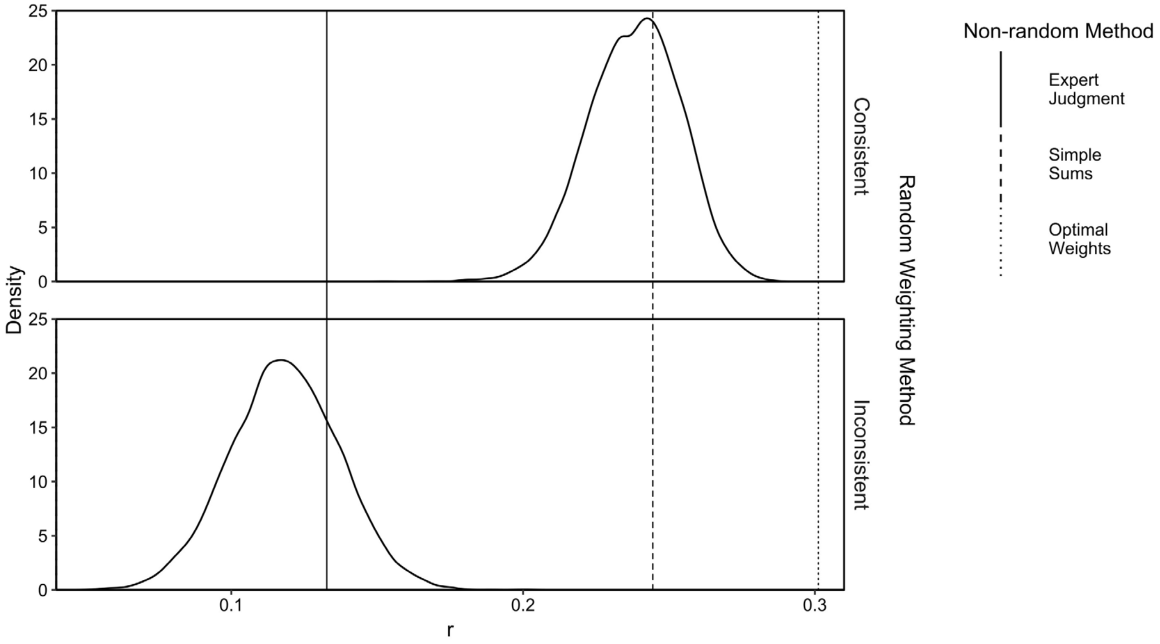


FIGURE 3.

Density distributions of validities (10,000 iterations each) at Company B, Sample 2 of predictor scores combined using random positive weights applied consistently (top plot) or inconsistently (bottom plot) across all candidates. Vertical lines are validities at Company A of non-random methods of data combination: expert judgment (solid line), unit weighting via simple sums (dashed line), and optimal weighting (dotted line).



in 22.22% of the iterations but never outperformed simple sums or optimal weights. 96.69% of the iterations for inconsistent weights were outperformed by all of the iterations for weights applied consistently.

DISCUSSION

Across the three samples, experts outperformed inconsistent weights 100%, 91.5%, and 77.8% of the time in predicting subsequent job performance ratings in Company A; Company B, Sample 1; and Company B, Sample 2, respectively. In turn, consistent weights outperformed experts 76.8%, 100%, and 100% of the time. These results indicate that experts do not make judgments completely inconsistently and are aware, to some extent, of what information is most valuable. However, their inconsistency in combining information does drastically damage their accuracy. This simulation study demonstrates that consistency in applying predictor weights is paramount to making accurate judgments.

It is striking that *mindless consistency* is enough to result in more accuracy than expert judgment. On average, random weights applied consistently resulted in better predictions than the assessors' own judgments, which parallels Dawes and Corrigan's (1974) earlier study of random weighting. In the Company A analyses, consistent use of random weights dominated the experts in the majority of cases. In the analyses for both Samples 1 and 2 at Company B, consistent use of random weights completely dominated the experts.

At this point, it is unclear what determines the extent to which consistent random weights will dominate over expert judgment, but it may be in part a function of the strength of the predictor–criterion relationships: Based on the predictive validity of the optimal weighting schemes, it is clear that scores on these assessment dimensions are better at predicting performance at Company B than at Company A. However, both Company B samples show 100% dominance of consistent random weights over expert judgment, but the optimal validity and the validity of expert judgment at Sample 2 are both lower than those at Sample 1. Another possibility is that dominance is dependent on the difference between the validities of expert judgment and optimal weighting as this difference is larger in both Company B samples compared to Company A. More research across a larger number of samples will be needed to decipher the mechanism underlying this dominance effect.

Differences in the validity of expert judgment across the three samples were fairly small. However, the dominance of expert judgment over inconsistent weighting was not the same. At Company A, expert judgment was completely better than inconsistent weights, but at both Company B samples, expert judgment was not always better than inconsistent weights. Possible explanations include that the assessors at Company A were simply more consistent or incorporating mechanical approaches to their judgment, or there may be differences in the variability of candidate characteristics between each sample that may impact how well an inconsistent weighting scheme would perform. Further research will be needed to determine organizational and individual differences that may influence the differences in validity between clinical expert judgment and mechanical methods of judgment. Nevertheless, it is troubling that expert assessors are not always better than inconsistent weighting, as it suggests that they do not necessarily understand what they are doing when combining information and evaluating candidates.

Ultimately, the finding that even random weights perform well when applied consistently suggests that consistency in applying predictor weights is more important than the weights themselves. Linear models are quite robust, and as long as the signs on the weights do not change (as is the case in the present study where all weights were positive), changes in weights are not expected to drastically impact their predictive power (Dawes, 1979). As Waller (2008) demonstrated with fungible weights, it is possible to derive an infinite number of alternate regression weighting schemes that yield a predictive validity almost as good as that of optimal weights (in multiple regression with three or more predictors). That being said, even though it is possible to generate a set of random weights that will perform very well when applied consistently, it can be difficult or impossible to tell how well that set of random weights will perform until the validation is conducted. In this simulation study, both optimal weights and unit weights via simple sums tend to perform better than random weights applied consistently. Practically speaking, if optimal weights are not known or cannot be approximated, it would be better to simply add up predictor scores instead of using an ill-defined weighting scheme.

The portion of inconsistent weights that produce composite scores that negatively predict job performance illustrate a primary concern of inconsistency in judgment. In these cases, configurations of random weights were generated such that there was some reversal of rank order. A candidate who would in reality have better job performance than another candidate is now predicted to have worse job performance. We know that this is not a reasonable prediction. When inconsistency is introduced into a decision system, options are no longer evaluated using comparable criteria, which has the result of making comparisons between options meaningless. Because judgmental consistency is difficult to accurately quantify until after the judgment has already been made, any inconsistency should be considered undesirable.

All that said, our intention here is to simply demonstrate the importance of consistency in making accurate judgments. Ultimately, our goal should be to improve consistency in human judgment, and we are reluctant to argue that clinical judgment should be completely replaced with unit weights. Hastie and Dawes (2001) stated that "whenever possible, human judges should be replaced by simple linear models" (p. 62-63). Like them, we think an emphasis should be put on the "whenever possible." Whereas mechanical methods are good for maximizing predictive power and minimizing costs and time, one concern is with face validity and reactions of both decision makers and the people affected by these decisions. Overall, people tend to perceive clinical methods of judgment to be more effective than mechanical methods, and mechanical methods have been described as unprofessional, impersonal, insufficient, inaccurate, unfair, and unethical (Diab et al., 2011; Eastwood et al., 2012). If useful decision aids are completely rejected or dismissed because of mechanical data combination, they cannot improve decision making. Adopting intermediate or blended approaches may help improve decision making while retaining user acceptance.

In practice, there are a number of ways in which the consistency of clinical judgments could be improved while maintaining the human aspect that many people strongly prefer. Here, we provide a sampling of methods for doing so. (a) Decision aids can be provided where judges enter weights that they themselves define into an algorithm. This would provide consistency and would likely improve prediction. Our results suggest that for one of the companies, experts were on average somewhat better than inconsistent weights, which means that they are aware of weighting strategies that improve prediction. (b) A "model of man" can be obtained where the judge's weighting policy is statistically estimated (Goldberg, 1970), and these weights can be entered into an algorithm or be provided back to the judge to allow him or her to better understand his or her own weighting policy. (c) Mechanical synthesis (Sawyer, 1966) can be used to retain clinical judgments, which are then mechanically combined with the original predictors into a final composite score. (d) Mechanical methods can be used to initially screen candidates, after which experts can then apply their own judgment to selecting among the top candidates. (e) To the extent that there is an effective underlying strategy, averaging the ratings of multiple experts would tend to reduce unreliability in weighting predictors and permit stronger correlations with relevant criteria.

Additional research on these methods includes identifying the methods that are most amenable to preserving positive reactions from everyone involved in a decision, evaluating the ability for each method to maintain predictive validity, and developing new methods of supporting expert judgment. This would be valuable from both practical and scientific standpoints. For the former, this would improve the utility of decision systems that involve expert judgment, and for the latter, better understanding of how to improve expert judgment will lead to a better understanding of the basic processes underlying judgment and decision making.

It is possible that the expert assessors could have had more information about the candidates beyond scores on these seven assessment dimensions, such as their performance on individual assessment activities, test profiles, and biographical information obtained from sources such as résumés and personal interaction. This could be viewed as an advantage that a human judge has over a mechanical method. Despite the possibility that the experts had this information available to them, they still performed worse than any mechanical method. Prior research has shown that although people tend to become more confident about their judgments with more information available, they are not always more accurate (Tsai et al., 2008).

Experts also potentially have insight into certain decisions that are not easily captured by a simple linear model. In a selection context, in some cases it may be more critical to identify the worst candidates than it is to identify the best candidates. Therefore, having expert insight into rare-occurring "red flags" not accounted for by the mechanical model (described by Meehl, 1954 as "broken-leg" cues) that signals whether a candidate possesses some fundamentally undesirable characteristic would provide crucial information in service of this goal. In evaluating such candidates where some red flag is highly diagnostic and overrides other information, the expert who is able to detect this red flag would be expected to provide a more accurate assessment than the mechanical model.

Yet, although experts may have insight, over the long run mechanical methods come out ahead. This is seen in the present study where, on average, expert judgment is nearly perfectly aligned with inconsistent weights in one case and only modestly outperforms inconsistent weights in another case. The issue is twofold. First, opportunities for insight to truly make a substantial difference are likely rare. In the case of red flags (or broken-leg cues), they are themselves defined as being rare occurrences. Therefore, good predictions using insight and bad predictions due to human error average out in the long run, and there are likely more opportunities to make errors than for insight to be important. Second, even with insight, people tend to overperceive and overgeneralize (Camerer & Johnson, 1991). Red flags tend to tell compelling stories, which leads to the inappropriate application of insight and to the neglect of relevant information and common sense (Highhouse, 2008). In light of these issues, a question is whether and how expert insight can be effectively captured and applied. Mechanical judgments can be highly accurate, but they are unable to account for any rare event that has not been included in the model or algorithm. Research into how expert insight can be effectively integrated with mechanical methods will hopefully further improve the predictive validity of our decision systems.

We note a couple limitations of our study. First, using a Monte Carlo simulation approach to generate random positive weights, we were able to examine the distributions of validities for the random-weighting methods. However, a more detailed analysis would have included distributions of validities of expert judgment across assessors in each dataset. Unfortunately, due to low within-assessor sample sizes, we could not be confident in the accuracy of validity estimates for individual assessors. We were therefore only able to examine assessors aggregated at each company and obtained a single average validity estimate across all assessors at each company. Future research with more substantial data should examine assessor-level differences through a multilevel modeling framework. With this, it would also be possible to combine multilevel models with the lens model to examine any differences in how each assessor assigns weights to the assessment dimensions and how these differences in turn impact the judgmental accuracy of each assessor (Kuncel, 2018).

Second, the effect sizes found in this study are likely local to the sample used for analysis. We do show that there are validity differences depending on the company for which assessors conducted assessments, and it is possible that a similar study using a different data source will also show effect sizes that depart from those found in this study. Additionally, it is possible that the one assessor could have conducted assessments for multiple companies, and if a high performing assessor conducts assessments for multiple companies, he or she could sway the aggregate predictive validity to be higher, and the opposite could be true for a low performing assessor. Unfortunately, we lack the identifying information in these archival datasets to be able to say for certain. That said, we expect that the substantive conclusions should hold, namely that mechanical methods of data combination on average outperform clinical judgment, and that judgmental consistency plays a large role in this because even random weights applied consistently often outperform clinical expert judgment a majority of the time.

In conclusion, no matter how strongly a set of predictors relate to the criterion, the predictive power of a decision system is dependent on how information is combined. Consistency in weighting predictors across all judgments heavily contributes to maximizing predictive validity. The bad news is that human judges and even experts are often inconsistent. The good news is that there are methods that can retain human judgment and potentially reduce human error and improve the consistency of human judgment while avoiding negative reactions toward the use of mechanical methods. Further research into these methods and continual development of new methods of improving judgmental consistency will ultimately improve our judgment and decision making processes, no matter the context.

REFERENCES

Brunswik, E. (1952). The conceptual framework of psychology. Chicago: University of Chicago Press.

Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In K. A. Ericsson & J. Smith (Eds.), Toward a general theory of expertise: Prospects and limits (pp. 195–217). Cambridge, UK: Cambridge University Press.

Dawes, R. (1979). The robust beauty of improper linear models in decision making. American Psychologist, 34, 571–582.

Dawes, R., & Corrigan, B. (1974). Linear models in decision making. Psychological Bulletin, 81, 95–106.

Diab, D. L., Pui, S., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in US and non-US samples. International Journal of Selection and Assessment, 19, 209-216.

Eastwood, J., Snook, B., & Luther, K. (2012). What people want from their professionals: Attitudes toward decision-making strategies. Journal of Behavioral Decision Making, 25, 458-568.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inference. 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, & Law, 2, 293–323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction. Psychological Assessment, 12, 19–30.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. Psychological Review, 62, 255–262.

Hastie, R., & Dawes, R. M. (2001). Rational choice in an uncertain world: The psychology of judgment and decision making. Thousand Oaks, CA: Sage.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. Industrial and Organizational Psychology: Perspectives on Science and Practice, 1, 333–342.

Jeanneret, R., & Silzer, R. (1998). An overview of individual psychological assessment. In R. Jeanneret & R. Silzer (Eds.), Individual psychological assessment. 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.

Kristof-Brown, A. L., Zimmerman, R. D., & Johnson, E. C. (2005). Consequences of individuals' fit at work: A meta-analysis of person–job, person–organization, person–group, and person–supervisor fit. Personnel Psychology, 58, 281-342.

Kuncel, N. R. (2018). Judgment and decision making in staffing research and practice. In D. S. Ones, N. R. Anderson, C. Viswesvaran, & H. K. Sinangil (Eds.), The SAGE handbook of industrial, work & organizational psychology (Vol. 2, pp. 474–487). London, UK: Sage.

Kuncel, N. R., & Highhouse, S. (2011). Complex predictions and assessor mystique. Industrial and Organizational Psychology: Perspectives on Science and Practice, 4, 302-306.

Kuncel, N. R., Klieger, D. M., Connelly B. S., & Ones, D.S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. Journal of Applied Psychology, 98, 1060-1072.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis, MN: University of Minnesota.

Morris, S., Daisley, R., Wheeler, M., & Boyer, P. (2015). A meta-analysis of the relationships between individual assessment and job performance. Journal of Applied Psychology, 100, 5-20.

Newman, D. A. (2014). Missing data: Five practical guidelines. Organizational Research Methods, 17, 372-411.

Putka, D. J. & Dorsey, D. W. (2019, April). A tour of I-O relevant AI/ML developments. Friday Seminar at the 34th Annual Society for Industrial and Organizational Psychology Conference, Washington, DC.

Ryan, A. M., & Sackett, P. R. (1987). A survey of individual assessment practices by I/O psychologists. Personnel Psychology, 40, 455–488.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. Psychological Bulletin, 66, 178–200.

Schenker, N., & Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. Computational Statistics & Data Analysis, 22(4), 425-446.

Schmitt, N., Noe, R. A., & Gottschalk, R. (1986). Using the lens model to magnify raters' consistency, matching, and shared bias. Academy of Management Journal, 29, 130-139.

Silzer, R., & Jeanneret, R. (2011). Individual psychological assessment: A practice and science in search of common ground. Industrial and Organiational Psychology, 4, 270-296.

Slaughter, J. E., & Kausel, E. E. (2013). Employee selection decisions. In S. Highhouse, R. S. Dalal, & E. Salas (Eds.). Judgment and decision making at work (pp 57–59). New York, NY: Routledge.

Tsai, C. I., Klayman, J., & Hastie, R. (2008). Effects of amount of information on judgment accuracy and confidence. Organizational Behavior and Human Decision Processes, 107, 97-105.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond and Hursch and by Hammond, Hursch and Todd. Psychological Review, 71, 528-530.

van Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). MICE: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45, 1-68.

Waller, N. G. (2008). Fungible weights in multiple regression. Psychometrika, 73, 691-703.

RECEIVED 05/09/19 ACCEPTED 03/23/20