Industrial and Organizational Psychology, **4** (2011), 302–306.
Copyright © 2011 Society for Industrial and Organizational Psychology. 1754-9426/11

Complex Predictions and Assessor Mystique

NATHAN R. KUNCEL
University of Minnesota

SCOTT HIGHHOUSE
Bowling Green State University

Silzer and Jeanneret (2011) made many declarations about incremental insight provided by individual assessment experts. We do not believe that sufficient evidence exists to support many of them. Our comment is focused on identifying those statements that are most inconsistent with the judgment and decision-making literature as well as providing recommendation for the kinds of evidence needed in the future. Although we vehemently disagree with many of the authors' assertions, and believe a great deal of data falsify their suppositions, we are encouraged by their call for additional and systematic research. We believe our ultimate goal is the same as the authors': to provide maximally valid, unbiased, accepted, and useful information to individuals and organizations.

Points of Contention

Throughout their article, Silzer and Jeanneret make assertions about insights and feats of data combination that are not supported by research. The research is clear that experts consistently make inferior predictions compared with simple equations, when predicting a variety of performance evaluations in both work and academic settings (Highhouse, 2002, 2008; Kuncel, Klieger, Connelly, & Ones, 2008). This pattern is wholly consistent with findings in a staggering range of fields and judgments (e.g., Grove, Zald, Lebow, Snitz, & Nelson, 2000). This broader literature, combined with meta-analytic work specifically in industrial–organizational (I–O) psychology (Kuncel et al., 2008), creates a grim picture for the practice of using expert judgment in both individual assessments and assessments centers. Below, we address some specific points of contention.

- *The assertion that assessors can synthesize interview information at the "data and construct" level simultaneously and that assessors can consider a myriad of variables for predicting both short-term and long-term outcomes.*

Research on decision making suggests that human judges simply cannot do the things suggested by Silzer and Jeanneret. Judges can handle around eight pieces of information (Cooksey, 1996), but no evidence supports their ability to deal with this many interactions let alone the vast list of variables (some multidimensional) listed in the target article. The problem is that decision-making research shows that experts do not make much use of

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

https://doi.org/10.1111/j.1754-9434.2011.01343.x Published online by Cambridge University Press

interactions. And, there are no data that we are aware of that support the notion that personnel selection experts use effective interactive rules. There is some evidence for nonlinear processing, but even this is limited to cognitively simplistic cut-off rules (conjunctive and disjunctive rules). Given the overall state of the evidence, claims of assessor cognitive acrobatics are not credible without strong empirical evidence.

- *The assertion that experts can complete both normative and ipsative interpretations for the same variables for the same assessee and that these are ''virtually impossible to complete in some mechanical or statistical manner.''*

Given that normative and ipsative have specific statistical definitions, both can be provided by mechanical methods. This assertion and others suggest that Silzer and Jeanneret believe in the ability of experts to conduct important configural or interactive interpretations of sets of variables. Contrary to lay intuition, interactions among abilities and traits are typically of little actual importance for making predictions. One type of interaction singled out in the target article is commonly referred to as the ''broken leg'' cue. It is an appealing notion that an expert could note a specific diagnostic indicator not captured by the selection system. Experts, however, tend to over identify the presence of such indicators and evaluate them as more important than they really are. This, in turn, results in poorer predictions and is likely to be a reason that feeding algorithmic judgments to experts fails to exceed algorithmic judgment used alone (Grove et al., 2000; Sawyer, 1966). Even if experts can use complex interactions, any meaningful higher-order interactions and configural rules could be incorporated into an equation. Whatever modest gains we might get over additive linear models would be better realized by an equation because human judges struggle with very simple interactions let alone the complex *Country Culture* × *Organizational Culture* × *Economic Environment* × *Supervisor* × *Organizational Challenge* × *Coworkers* × *Individual* × *Time* interactions discussed in their article.

- *The assertion that candidates were ''easier to profile'' in the past but are more difficult to profile now.*

We are aware of no empirical evidence indicating that assessees have become easier to profile let alone that experts can manage such an issue or have adapted to such changes. Further, reweighting sets of predictors to adapt is unnecessary. First, simple unit weights consistently beat experts in decision making. This means that no validation study is needed to derive regression weights—one may simply standardize job-related predictors and add them together to make sound predictions. Given the current understanding of job performance predictors, they also require no updating over time. Moreover, the structure of performance and its determinants would need to change radically for unit weights to become invalid—including variables reversing validity sign or dropping to zero validity. Such events are highly implausible. The literature to date indicates that predictors have remained valid and at roughly the same level for decades for all major classes of predictors. No substantial evidence exists to support the existence of this problem. No evidence demonstrates that experts can and have done so for predicting job performance or that they can do so better than a simple equation.

- *The assertion that assessors can both adapt their evaluations to ''changing organizational contextual variables and job demands'' and that these adaptations improve predictive accuracy.*

We are aware of no evidence supporting the contention that experts can reliably detect shifting population weights. In fact, experts are often resistant to change,

https://doi.org/10.1111/j.1754-9434.2011.01343.x Published online by Cambridge University Press

even when presented with contradictory evidence. Even if some of these abilities do exist in the real world, there is a critical issue here that goes unaddressed by Silzer and Jeanneret: Experts must perform with enough consistency and effectiveness to overcome the numerous and serious problems associated with expert prediction, including the issues of inconsistency, illusory correlation, and biased judgment. That is, individual assessment experts not only have to have special insights, but the insights must also compensate for very normal human weaknesses.

Research Needs

Silzer and Jeanneret argue for the presence of specific superior experts in the field. It is important to note there is a reasonable database on expert judgment predicting human performance in work and academic settings. The literature on expert judgment is based on *real experts charging clients for their services*, and the data show that real psychologists are being bested by simple addition. Experts who are superior would need to put forth very strong empirical evidence supporting their performance because the data across many fields suggests that this is not possible. Therefore, the real issue is not that some individuals are relatively better than others, the issue is whether they are better than alternatives (equations).

The only space remaining is to claim that the contribution of experts is based on the entirely unique features of the situation that no model can capture and that the benefits of these idiographic analyses overcome the costs. Such data could be collected in which individual assessment experts make very specific and verifiable predictions about what will happen in the future for one unique individual in one unique situation. We note that such research has been done in other fields and the conclusions have not remotely supported expert judgment (e.g., Tetlock, 2005). Tetlock conducted a study of 27,451 predictions from experts and ''dilettante'' about the future in the area of political and economic analysis. Experts were little better than dilettante, and Tetlock likened the experts to ''dart throwing chimps'' (p. 80). Although some experts appeared to be better than others, they were *all* inferior to crude algorithms, and better performance had nothing to do with experience. Perhaps experts in I–O would fare better, but strong empirical evidence would be needed. We are encouraged by Silzer and Jeanneret's call for more research in the field because it is sorely needed.

The study by Silzer (1984) is cited as a main piece of evidence within the I–O literature indicating that individual experts can be better than mechanical methods. First it is important to note that the data from Silzer (1984) do not fit the Silzer and Jeanneret preferred definition of individual assessments because the assessors in the study did not have actual contact with the applicants. Second, the study is based on a very small number of assessors. Data from an individual assessment study that fit the preferred definition (Kuncel, 1999) revealed only one assessor who matched, but did not best, the validity of the equation. A policy capturing study was begun to try to understand why the assessor was as successful as the model. This effort was abandoned, however, because the expert stated that he simply ''added them up'' to form a final evaluation.

More important than any two studies is that we live in a post-Schmidt and Hunter (1977) world. Identification of specific studies that support or refute a position is less important than the overall message of the literature as a whole. The burden of proof for demonstrating why the effects in those studies are meaningfully different from the overall sweep of the literature lies with the critics. In the overall sweep of the literature, individual experts who are superior to equations cannot be reliably identified for these kinds of tasks. For example, over multi year spans of time, individual mutual fund managers may outperform a simple market average, but that superiority is not maintained (e.g., Malkiel,

https://doi.org/10.1111/j.1754-9434.2011.01343.x Published online by Cambridge University Press

1995) indicating the effects of chance. Even when better-than-average experts are identified, they are not as good as equations (Sherden, 1998; Tetlock, 2005).

Toward Evidence-Based Practice

Silzer and Jeanneret disagree with idea that most I–O psychologists who do individual assessment take a holistic or ''whole person'' approach to the practice. Prien, Shippmann, and Prien (2003), however, said that the holistic approach is a ''hallmark of individual assessment practice.'' In their classic book on individual assessment, the authors themselves said, ''Assessors need to understand the interaction of beliefs, values, and behavior in order to understand the individual in a more holistic way'' (Silzer & Jeanneret, 1998).

The holistic assessor emphasizes the uniqueness of the individual and downplays nomothetic generalizations. Ruscio (2003) noted that, ''Holistic judgments are premised on the notion that interactions among all of the information must be taken into account to properly contextualize data gathered in a realm where everything can influence everything else.'' Silzer and Jeanneret reflect this school of thought when they discuss the ''central role'' of the psychologist and the critical role of psychological judgment in assessment. As we have outlined in this article, however, the holistic school of thought whether producing a single decision or a multi faceted evaluation is not consistent with the evidence base.

Individual assessment is an area of practice where one's livelihood can sometimes come into conflict with objective scientific discovery. We believe that, all else equal, the use of subjective expert judgments in individual assessment is inconsistent with evidence-based practice. We encourage the field to move away from expert judgment for data combination and pursue research on methods that simultaneously attempt to maximize both end-user adoption and predictive power (Kuncel, 2008).

Conclusion

There is little room left for assertions of incremental insight provided by individual assessment experts, given the clear evidence that (a) experts are inferior at data integration, (b) expert judgments are explained by simple decision rules, (c) experts struggle to consider even simple configural rules, (d) mechanical methods can effortlessly include both nonlinear and interactive components, (e) but don't even need them to beat human judges. Simple linear models alone capture nearly all of the variance in the decisions of expert judges, and those models beat the judge on which they are based. These facts effectively rule out the notion that experts actually do the things claimed by Silzer and Jeanneret.

We are not bothered that many human technologies outperform us in many ways. Our cars move faster than we can run, aircraft piloting programs make unmanageable stealth planes manageable, and computers beat world-class *Jeopardy* contestants. Is it really so bad that formulas outperform psychologists in combining information to make predictions? Providing an overall recommendation is the norm in individual psychological assessments (Ryan & Sackett, 1987), and the literature is clear that they are of lower quality when made by an expert. Nevertheless, matching the performance of an equation *is truly* within the information processing capabilities of most experts. Mental arithmetic can be learned.

References

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications.* San Diego, CA: Academic Press.

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

Highhouse, S. (2002). Assessing the candidate as a whole: A historical and critical analysis of individual psychological assessment for personnel decision making. *Personnel Psychology, 55*, 363–396.

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

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

https://doi.org/10.1111/j.1754-9434.2011.01343.x Published online by Cambridge University Press

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.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2008). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. In I. L. Kwaske (Chair). *Individual assessment: Does the research support the practice?* Symposium presented at the 23rd Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Malkiel, B. G. (1995). Returns from investing in equity mutual funds 1971 to 1991. *The Journal of Finance, 1*, 549–572.

Prien, E. P., Schippmann, J. S., & Prien, K. O. (2003). *Individual assessment: As practiced in industry and consulting.* Mahwah, NJ: Erlbaum.

Ruscio, J. (2003). Holistic judgment in clinical practice: Utility or futility? *The Scientific Review of Mental Health Practice,* 2. Retrieved from http://www.srmhp.org/0201/holistic.html.

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.

Schmidt, F. L., & Hunter, J. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529–540.

Sherden, W. A. (1998). *The fortune sellers: The big business of buying and selling predictions.* New York, NY: John Wiley and Sons.

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

Silzer, R., & Jeanneret, R. (1998). Anticipating the future: Assessment strategies for tomorrow. In R. Jeanneret & R. Silzer (Eds.), *Individual psychological assessment: Predicting behavior in organizational settings* (pp. 445–477). San Francisco, CA: Jossey-Bass.

Silzer, R., & Jeanneret, R. (2011). Individual psychological assessment: A practice and science in search of common ground. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 270–296.

Tetlock, P. (2005). *Expert political judgment.* Princeton, NJ: Princeton University Press.

https://doi.org/10.1111/j.1754-9434.2011.01343.x Published online by Cambridge University Press