Personnel Assessment and Decisions

Volume 6 | Issue 3 Article 2

2020

The Machines Aren't Taking Over (Yet): An Empirical Comparison of Traditional, Profiling, and Machine Learning Approaches to Criterion-Related Validation

Kristin S. Allen
SHL

Mathijs Affourtit
SHL

Craig M. Reddock

Recommended Citation

Allen, Kristin S.; Affourtit, Mathijs; and Reddock, Craig M. (2020) "The Machines Aren't Taking Over (Yet): An Empirical Comparison of Traditional, Profiling, and Machine Learning Approaches to Criterion-Related Validation," *Personnel Assessment and Decisions*: Number 6 : Iss. 3 , Article 2.
DOI: https://doi.org/10.25035/pad.2020.03.002
Available at: https://scholarworks.bgsu.edu/pad/vol6/iss3/2

The Machines Aren't Taking Over (Yet): An Empirical Comparison of Traditional, Profiling, and Machine Learning Approaches to Criterion-Related Validation

Kristin S. Allen[1], Mathijs Affourtit[1], and Craig M. Reddock

1. SHL

KEYWORDS

validation, machine learning, profiling

ABSTRACT

Criterion-related validation (CRV) studies are used to demonstrate the effectiveness of selection procedures. However, traditional CRV studies require significant investment of time and resources, as well as large sample sizes, which often create practical challenges. New techniques, which use machine learning to develop classification models from limited amounts of data, have emerged as a more efficient alternative. This study empirically investigates the effectiveness of traditional CRV with a variety of profiling approaches and machine learning techniques using repeated cross-validation. Results show that the traditional approach generally performs best both in terms of predicting performance and larger group differences between candidates identified as top or non-top performers. In addition to empirical effectiveness, other practical implications are discussed.

There is pressure to demonstrate value and return on investment when administering prehire assessments whether organizations are fighting to hire top talent (Sullivan, 2014; The Talent Board, 2019) or in times of economic downturn when budgets are at risk (Fernandez-Araoz, 2020). Regardless of the state of the economy, industrial and organizational psychology (I-O) practitioners must balance the needs of the organization with the rigor required to ensure that assessments used to inform hiring decisions are job relevant, fair, and defensible. When developing and implementing selection procedures, I-O practitioners are guided by legal and professional guidelines. For example, in the United States legal standards (EEOC Guidelines, 1978) and professional best practices (Society for Industrial and Organizational Psychology [SIOP Principles], 2018) guide practitioner decisions. Likewise, practitioners in other regions of the world follow local legal guidelines and professional best practices.

Typically, criterion-related validation (CRV) studies are performed to establish evidence of validity (Gatewood et al., 2008) to support the use of test scores from an assessment to make hiring decisions. Best practices for conducting CRV studies have remained consistent for many years. However, organizations are often reluctant or unable to invest the resources necessary to complete a proper criterion-related validation study (Johnson et al., 2010; Van Iddekinge & Ployhart, 2008), unless they are legally required to do so, leaving I-O practitioners searching for viable and legally defensible alternatives.

Techniques have emerged that enable practitioners to validate assessments in less than ideal circumstances (e.g., limited data, small sample sizes; Putka et al., 2018). Such methods utilize concepts typical of other disciplines like computer science, artificial intelligence, and machine learning (Gonzalez et al., 2019; Putka et al., 2018); can be used to measure personality (Alexander et al., 2020); and can predict turnover and performance (Sajjadiani et al., 2019). Some of these methods use limited amounts of data to develop classification (or profiling) models based on machine learning techniques. Although these techniques allow for more efficient test validation, it is critical to determine

Corresponding author:
Kristin S. Allen
Email: Kristin.Allen@shl.com

their effectiveness. This study will empirically evaluate the effectiveness of several approaches to creating predictive models, extending previous research by comparing machine learning based profiling techniques as well as a broader set of machine learning algorithms to a traditional validation study methodology.

Profiling Methods Using Machine Learning

Traditional validation studies typically involve conducting a job analysis, administering a battery of job-related assessments to candidates or incumbents, collecting job performance data, and examining the relationship between the predictor and criterion as a way to establish evidence of validity (Gatewood et al., 2008). Although these methods undoubtedly comply with legal, professional, and best practice guidelines for validating assessments such as the SIOP Principles (2018) and the EEOC Uniform Guidelines (1978), they also require significant organizational resources such as collecting assessment data from sufficiently large samples, collecting performance ratings, and/or collecting objective measures of performance (e.g., sales dollars). The time and resource investments necessary for a proper validation study are substantial.

In recent years, profiling approaches have been employed as an alternative to traditional CRV studies. In fact, research shows that 62% of consulting organizations implement some form of profile matching (Kulas, 2013), yet there is limited evidence regarding the efficacy of such practices. Profiling methods vary, as described in Table 1, but all are based on the premise of creating a profile of what a top performer looks like and applying that profile to identify top performers from a pool of candidates. This method requires the organization to identify a subset of high performing employees. It might also include a subset of low performing employees with the goal of identifying those who are likely to underperform. The way in which top and bottom performers are identified varies, ranging from subjective judgments made by managers to more objective job-related performance metrics. Once the subset of performers are identified, the incumbents complete one or more assessments. The assessment scores are used to create scoring algorithms, which are then used to evaluate a candidate's likelihood of success. The scoring algorithm indicates how closely the candidate matches the profile of a top performer. However, these basic profiling techniques have been found to offer lower observed estimates of predictive validity when compared to a more traditional approach using linear regression (Kulas, 2013).

Profiling methods essentially dichotomize the available sample, and evidence supporting this practice is mixed. Kelley (1939) concluded that taking 27% of participants from the extremes was optimal for evaluating item performance. More recently, Hunter and Schmidt (1990) suggest corrections to overcome the attenuating effect of dichotomization on observed correlations, whereas other researchers (MacCallum et al., 2002) assert that dichotomization is rarely defensible.

In an effort to understand how profiling approaches are typically implemented by I-O practitioners, we reached out to our networks and interviewed practitioners who have current or previous experience working in organizations that employ profiling techniques for personnel selection. Participants were informed that their names or affiliations would not be shared and the information they provided would only be used to help inform the research design to increase the practical relevance of this study. Based on these interviews, it became clear that many organizations utilize a technique that involves creating a profile based on a small sample of top and bottom performers. For this reason, the focus of this paper is on small sample profiling techniques, which may not be as effective as large sample techniques that make use of very large data sets.

When compared to traditional validation study methods, profiling approaches are appealing because they require fewer organizational resources. There is no established standard for how many performers are needed to establish a profile, so organizations are left to make their own judgments about the sample size required. According to Schmidt et al. (1976), the average sample size for a validation study is 68, which gives researchers only a 50% chance of detecting evidence of validity if it exists. Considering the advances in technology over the last 40+ years that have enabled more efficient data collection, this figure is likely an underestimate. In a more recent study, Bosco et al.'s (2015) meta-analysis reported sample sizes of 190, 202, 158, and 200 for relationships between performance and attitudes, knowledge, skills and abilities, psychological characteristics, and objective person characteristics, respectively. If a similar number of top performers is needed to achieve the statistical power to establish a stable algorithm, then organizations using smaller samples assume significant risk creating profiles. However, advancements in machine learning, described in Table 1, may offer an option for more efficiently conducting CRV studies. For example, techniques such as support vector machines (SVM) are trained to distinguish between the two groups (i.e., top performers and non-top performers) and used to create a machine learning model that identifies top performers.

A broader range of machine learning techniques are also gaining popularity. In the recent SIOP Top 10 Workplace Trends survey (SIOP, 2020), nearly 1,000 psychologists voted "artificial intelligence and machine learning" as the #1 trend that organizations are likely to face in 2020. This rising interest is also reflected in the popularity of the annual SIOP machine learning competition. In an effort to maximize the relevance and usefulness of this study to I-O

TABLE 1.
Description of Validation Approaches

Approach		Description
Traditional	Predictive Scales	Contains only the predictive scales method where multiple scales are administered and the most predictive scales are selected for inclusion in the final score.
Profiling	Support Vector Machine	An SVM model was trained to classify respondents as either top of bottom performers (Chang & Lin, 2018; Schölkopf et al., 2000), a grid search was applied during the training phase to ensure effective modelling parameters were used. Support vector machine models construct hyperplanes to divide observations into different categories while maximizing the distance to the nearest data point.
	Mean and SD Scaled Distance	The distance between a respondent's scores and top performers scores are calculated and corrected for differences in variance by dividing by the standard deviation values of top performer's scores. Next an overall score is calculated across all scores by calculating the Euclidean distance from the top performer's scores. Respondents with a smaller than average distance are classified as top performers (Nunnally & Bernstein, 1994).
	Mean and SD Band	Score bands are created based on the mean and standard deviation values of the scores of top performances. Future respondent's scores have to fall into these bands on each of the scales for them to be classified as top performers. For this study a band of mean +/-1.5 SD was used (Kulas, 2013).
	Random Forest Classifier	A random forest classifier was trained to classify respondents as either top or bottom performer (Breiman, 2001). A grid search was applied to ensure appropriate parameters were used. Random forest methods construct multiple decision trees models, using bootstrapped samples drawn from the data set to create an ensemble model.
	Profile Similarity (r)	The pattern between a respondent's score is compared against the average of the scores of top performers by calculating the correlation between these two sets of scores (Kulas, 2013).
Machine Learning	Regression	A regression equation is fit to the data to optimally predict performance.
	Ridge Regression	In addition to a regression equation fit to the data, a regularization factor is included to prevent overfitting to the training data (Friedman et al., 2010). A grid search was applied to ensure appropriate parameters were used.
	Random Forest Regression	This approach aims to predict a respondent's future job performance by fitting a random forest regression model (Breiman, 2001). A grid search was applied to ensure appropriate parameters were used.

practitioners, this study also includes a comparison of the efficacy of select machine learning techniques for validating assessments.

Although these new techniques may hold promise for increasing efficiency, the impact on the accuracy of predictive models remains unclear. In real-world situations validation studies are complex and involve careful balancing of the tension between best practices and practical considerations. The purpose of the present study is to empirically evaluate different approaches with regards to criterion-related validation and to inform I-O practitioners about the merits and limitations of each.

METHOD

Sample

The data were collected from job incumbents in a telecommunications company located in the United States as part of a concurrent validation study. Twenty behavioral scales from a self-report behavioral competency assessment were selected for inclusion in the validation study, based on the results of a job analysis. Supervisors of the participating incumbents completed a research-based job performance rating (JPR) survey on performance areas that directly aligned with the behavioral competencies measured in the assessment, as well as five ratings of global performance such as "How would you rate the employee's overall job performance?" and "How would you rate the overall match between each employee's ability and the job requirements?". The items were presented on a 5-point scale (ranging from 1 = *below average* to 5 = *one of the best*). Supervisors were asked to rate all participating direct reports on each item on the same page of the survey. This comparative rating scale was used to encourage differentiation of ratings and reduce rating biases (e.g., leniency and severity), as it prompts the supervisor to consider how direct reports rank against each other in each of the performance areas. The number of direct reports rated by each supervisor ranged from 6 to 12 direct reports (9.3 on average). The five global performance items were combined into a composite that will be used in this study (α = .90), which will be referred to as overall job performance ratings. Participating supervisors were informed that the ratings would be kept strictly confidential and not shared with anyone within their organization. Additionally, scores rated by leadership on the overall quality of the individual's messages when interacting with customers were included. This indicator of performance will be referred to as quality scores. Reliability was calculated using ICC 1 with a one-way random effect model, ICC = .83 ($F(183,368) = 5.836$, $p < .001$).

The original sample contained 208 individuals with assessment data and 232 with overall job performance ratings. All records without excessive missing data were initially identified as viable and then were dropped (cumulatively) if they did not meet the following criteria: When asked how familiar the rater is with the employee's job performance, the rater responded at least "fairly well" and when asked how often the rater has the opportunity to observe the employee's performance, the rater responded at least "once per month," resulting in a reduction of 3% of the available sample. The final cleaned and matched data set consisted of 202 and 169 cases with overall job performance ratings and quality scores, respectively.

Analyses

Validation studies commonly test more content than is expected to be used in the final assessment battery, allowing the organization the flexibility of choosing the most predictive scales while taking into account other considerations (e.g., testing time, candidate experience, etc.). In practice, this approach is informed by a job analysis and conceptual review of the scales being considered to ensure there is both empirical and conceptual support for the chosen measures. The same approach was used here, taking the six most predictive behavioral scales based on the observed estimates of composite predictive validity. Although the estimates of composite predictive validity may have increased by selecting a combination of more than six scales, six were chosen to maximize estimates of validity while minimizing testing time to reflect how this strategy is applied in practice.

The comparative observed estimates of predictive validity for the various methods was established using repeated cross-validation, which is sometime referred to as Monte Carlo cross-validation (Borra & Di Ciappo, 2010; Kuhn & Johnson, 2013). In repeated cross-validation, the dataset is repeatedly split into a training and test set. Repeated cross-validation produces an estimate of the expected validity, while reducing sampling bias that would result from analyzing a single sample. Predictive models are created using the training set, and the effectiveness of these models is evaluated on the test set. For this study 70% of the data was used in the training set and 30% in the test set. This process was repeated 100 times using different subsamples.

The three sets of validation techniques described in Table 1 are compared. First, the traditional validation approach, which contains only the predictive scales method where multiple scales are administered and the most predictive scales are selected for inclusion in the final score. Second, a set of profiling approaches that use data from performers classified as top and bottom performers. In an effort to mimic the various ways by which top and bottom performers are identified in practice, three different sampling approaches were used as described in Table 2. Although using 10 candidates in the top and bottom categories is remarkably small from a statistical perspective, based on our research this is representative of what is being used for small sample profiling practices. Finally, a set of machine learning techniques was evaluated for effectiveness in making predictions about work performance, as described in Table 1.

For each repetition, after selecting the predictive scales using the training sample and training the classification models, the models were used to make predictions of job performance on the participants in the test sample. To most closely resemble the way this is done in practice, the traditional method was restricted to six scales that were selected in the training phase, while the other methods utilized all scales. Two criteria were used to evaluate the effectiveness of each method. First, the correlation between the predictor and the performance composite was calculated. When

TABLE 2.
Sampling Techniques

Sampling technique	Description of methodology
Full sample	• Taking the top 10 and bottom 10 candidates from the full sample based on their global performance composite scores. • Simulates the scenario where elaborate job performance data are available for a large pool of job incumbents. • To implement this in practice a full-scale job performance rating exercise would need to be conducted, potentially attenuating the efficiency benefit of using a profiling approach.
Subsample	• Taking a subset of 20 candidates who were randomly selected from the training set. • These candidates were then divided equally into 10 top and 10 bottom performers based on the global performance composite scores. • Simulates the scenario where a limited number of job incumbents are available for participation in the validation study.
Overall rating sample	• Randomly selecting 10 top performers from the subset of candidates who had received a score of 4 or a 5 (on a 5-point scale) on a single item measuring overall job performance. • Ten bottom performers were randomly selected from the subset of candidates who had received a score of 1 or 2 on the same item. • Simulates the scenario where supervisors or managers make subjective judgments on who falls into the top or bottom categories without doing a research-based job performance rating survey with carefully defined job-related criteria.

measures are obtained from a sample where the variance is higher than the population (as would be the case with profiling), then the observed correlation can be inflated (Nunnally & Bernstein 1994). For this reason, the performance of the profiling methods is evaluated on the entire testing sample. In I-O psychology, correlations between the predictor and criteria are generally the most common measure to demonstrate evidence of criterion-related validity (Gatewood et al., 2008). However, because classification models produce dichotomous scores that can only assume two values, correlations are less appropriate to evaluate the estimates of validity of classification models. Therefore, in addition to the correlation coefficient, the effect size of group differences between the top and bottom performers were calculated. Effect sizes were calculated using the d statistic, the standardized mean score difference between groups. A *d* value of 0.2 is considered small, 0.5 is considered medium, and 0.8 is considered large (Cohen, 1988). Larger effect sizes indicate a larger difference in the performance of the two groups, with a positive effect size indicating better performance by those identified as top performers. To calculate an effect size for the traditional scale selection and the nonclassification machine learning techniques approach, all participants in the testing sample were divided into top and bottom performers using the predicted scores.

RESULTS

The mean and standard deviation values of aggregated correlations and effect sizes across repetitions by approach are shown in Table 3 for predicting overall job performance ratings and Table 4 for predicting quality scores. The traditional (predictive scales) method shows a correlation of .159 with overall job performance ratings, whereas the correlations for profiling approaches range from -.077 to .088, and the machine learning techniques range from .055 to .173. Similarly, when predicting the quality scores, the correlations were .208 (predictive scales), ranging from -.107 to .176 (profiling approaches), and ranging from .139 to .233 (machine learning techniques).

When predicting overall job performance ratings, the group difference (effect size between participants identified as top and bottom performers) was .290 for the predictive scales method, -.155 to .206 for profiling approaches, and .061 to .236 for the machine learning techniques. When predicting the quality scores, the effect size between participants identified as top and bottom performers was the largest with a correlation of .347 (predictive scales), ranging from -.190 to .279 (profiling approaches), and ranging from .223 to .321 (machine learning techniques).

Across these findings, results converge to suggest that

TABLE 3.

Predicting Overall Job Performance Ratings (100 repetitions)

Approach		Sampling method	Correlation			Effect size *(d)*		
			Mean	*SD*	*N*	Mean	*SD*	*N*
Traditional	Predictive scales	N/A	0.159	0.117	100	0.290	0.252	100
Profiling	Support Vector Machine	Full	0.088	0.135	100	0.125	0.253	100
		Sub	-0.004	0.155	100	-0.020	0.268	100
		Overall	0.020	0.134	100	0.013	0.249	100
	Mean and SD Scaled Distance	Full	-0.077	0.112	100	-0.155	0.207	100
		Sub	-0.043	0.114	100	-0.057	0.203	100
		Overall	-0.052	0.108	100	-0.064	0.228	100
	Mean and SD Band	Full	0.072	0.131	100	0.206	0.355	99
		Sub	0.033	0.110	100	0.110	0.380	94
		Overall	0.040	0.111	100	0.133	0.453	95
	Random Forest Classifier	Full	0.039	0.112	100	0.079	0.231	100
		Sub	-0.002	0.120	100	-0.004	0.246	100
		Overall	-0.030	0.126	100	-0.061	0.258	100
	Profile Similarity (r)	Full	0.078	0.135	100	0.102	0.252	100
		Sub	0.040	0.132	100	0.035	0.213	100
		Overall	0.055	0.128	100	0.026	0.238	100
Machine learning	Regression	N/A	0.173	0.104	100	0.236	0.245	100
	Ridge Regression	N/A	0.116	0.103	100	0.163	0.238	100
	Random Forest Regression	N/A	0.055	0.112	100	0.061	0.224	100

Note. In some cases, the Mean and SD Method did not identify sufficient top performers (two are required) in the test sample to compute an effect size. In these cases, the number of repetitions is less than 100.

TABLE 4.

Predicting Quality Scores (100 Repetitions)

Approach		Sampling method	Correlation			Effect size *(d)*		
			Mean	*SD*	*N*	Mean	*SD*	*N*
Traditional	Predictive scales	N/A	0.159	0.117	100	0.290	0.252	100
Profiling	Support Vector Machine	Full	0.176	0.118	100	0.279	0.236	100
		Sub	0.003	0.179	100	0.031	0.305	100
		Overall	0.081	0.121	100	0.128	0.233	100
	Mean and SD Scaled Distance	Full	-0.107	0.118	100	-0.166	0.252	100
		Sub	-0.068	0.123	100	-0.097	0.247	100
		Overall	-0.104	0.113	100	-0.190	0.269	100
	Mean and SD Band	Full	0.022	0.141	96	0.072	0.567	80
		Sub	0.036	0.155	98	0.124	0.556	96
		Overall	0.031	0.118	97	0.103	0.501	85
	Random Forest Classifier	Full	0.115	0.120	100	0.236	0.250	100
		Sub	0.003	0.159	100	0.006	0.330	100
		Overall	0.037	0.129	100	0.076	0.267	100
	Profile Similarity (r)	Full	0.145	0.130	100	0.268	0.264	100
		Sub	0.062	0.135	100	0.077	0.301	100
		Overall	0.098	0.131	100	0.229	0.292	100
Machine learning	Regression	N/A	0.173	0.104	100	0.236	0.245	100
	Ridge Regression	N/A	0.116	0.103	100	0.163	0.238	100
	Random Forest Regression	N/A	0.055	0.112	100	0.061	0.224	100

Note. In some cases, the Mean and SD Method did not identify sufficient top performers (two are required) in the test sample to compute an effect size. In these cases, the number of repetitions is less than 100.

the predictive scales approach consistently outperforms the profiling approaches. To evaluate whether differences between methods were significant, significance testing was performed on the resulting observed estimates of predictive validity coefficients. The predictive scales method was used as a baseline to which other methods were compared. Given that the performance criteria were computed using the same testing sample across methods within each iteration, a pairwise t-test was conducted after applying Fisher's r to Z transformation (Fisher, 1915) to the correlations. All differences between the predictive scales approach and profiling approaches were found to be significant ($p < 0.05$).

Furthermore, the predictive scales approach tends to outperform machine learning approaches, with the exception of regression, which produced higher correlation coefficients for overall job performance ratings. However, these differences were not significant. Finally, ridge regression produced significantly higher correlations with the quality scores but showed smaller correlations with the overall job performance ratings and smaller group differences across both criteria.

DISCUSSION

The results showed that the predictive scales technique outperformed the profiling approaches in terms of estimating the prediction of overall job performance ratings and the quality scores. When comparing the methods, both correlations coefficients and effect sizes were significantly different. This performance difference could be explained by the fact that profiling methods use limited numbers of observations that, combined with the dichotomization of the data, leads to loss of predictive information that could be obtained from the sample.

The traditional predictive scales technique also outperformed the machine learning techniques in predicting overall job performance ratings, with the exception of regression (where the validity coefficient was greater, but the difference was not statistically significant). Machine learning techniques have the ability to pick up on subtle trends in the data, however those may be sample specific. Traditional methods represent a balanced approach where relevant information is retained without amplifying spurious relationships. Our findings are consistent with previous research (Kulas, 2013) where the profiling techniques exhibited significantly lower validities, and regression techniques performed best. For predicting the quality scores, ridge regression demonstrated a significantly greater validity coefficient than the traditional predictive scales technique. The predictive scales technique outperformed all other methods.

When considering sampling approaches for the profiling techniques, the full sample technique was the most predictive approach. Validities for the other two profiling techniques were near zero and sometimes in the negative direction. When evaluating group differences in performance between the groups of top and bottom performers, the full sample profiling approach was generally the most effective of the various profiling approaches but still showed a negligible effect size, whereas the traditional predictive scale approach showed an effect size in the small to medium range. It should be noted that the best effect size for the profiling approaches was obtained in the most favorable scenario (full sample) where research-based job performance ratings for a large pool of job incumbents were available, which is not likely how performers are typically identified when profiling approaches are employed.

Important Considerations

Legal defensibility. Beyond the differences in accuracy of hiring decisions, there are other important considerations for practitioners to keep in mind when selecting a validation method. The first is legal defensibility. Typically, profiling approaches are not preceded with job analysis, which is a critical step in determining the job relevance of an assessment to comply with legal and best practice guidelines. Although a profiling approach seeks to replicate the behaviors or characteristics exhibited by top performers, the characteristics identified in the profile of top performers are not necessarily job related, yet this information serves as the basis for evaluating job candidates. Profiling takes a person-oriented approach, instead of a job-oriented approach, where the candidate is assessed against the requirements of the role. Any approach used in practice should be substantiated with job analysis evidence to demonstrate the job relatedness of the attributes measured by the assessment and the relevance of the criteria against which the assessment is validated to ensure compliance with legal requirements.

Further, machine learning approaches often use complicated black box algorithms that make the outcomes challenging to interpret and can present legal challenges. The most transparent techniques are regression based, as coefficients in these techniques can be interpreted as importance weights. However, weights are sometimes assigned in a manner that seems counter intuitive such as when there are suppressor effects (Paulhus et al., 2004). Therefore, practitioners should review the weights assigned by regression models to ensure that they are sensible.

Sample size. Although the thought of replicating top performers in new hires may be appealing to hiring managers, it's important to keep in mind that profiling approaches typically use only a small amount of data (e.g., assessment scores on 10 top performers). It is certainly more convenient to collect data from a small subset of people rather than a full sample of 200+ incumbents and their managers. However, a scoring algorithm developed on a small group of individuals is susceptible to sampling error and may not be generalizable to a group of future job candidates. If a profiling approach is used, ongoing validation should be

employed to confirm that those who have been selected for hire are indeed likely to be top performers in the job role. Future research should examine the impact of sample size on the accuracy of profiling and machine learning models to determine what, if any, sample size might approximate the results of a traditional approach.

Machine learning approaches require larger samples than the traditional or profiling methods, and it is typically very challenging to build predictive models using small samples as common place in validation studies. Initial models should be cross-validated as part of development, and ongoing validation efforts should be employed.

Fairness. Additionally, using the profiles of existing incumbents who are strong performers to select job candidates limits the maximum potential for high performance, assuming that the existing top performers are the best possible performers for the job role. Particularly with the rapidly changing nature of the world of work, the demands of jobs are evolving, and candidates may possess attributes that would make them even more successful in the job role than the existing top performing incumbents. When employing a profiling approach, the potential for nonincumbent job candidates to be successful in new and different ways can be overlooked, if the selection system is fixated on incumbents rated as top performers based on possibly outdated definitions of successful performance criteria.

When only considering the candidate's score distance from the top performer profile score, important information about the candidate may be overlooked. This also raises an ethical question of whether it is fair to make hiring decisions based on how well the candidate matches the profile of the top performers rather than selecting the candidates who performed best on the assessment.

Diversity. Finally, research shows that teams consisting of individuals with diverse characteristics and thinking styles will maximize high performance (Gartner, 2018). By using a profile of top performers to select new hires, the organization may be restricting diversity with efforts to hire more individuals that are highly similar to existing top performers. This issue is exacerbated when top performers are identified subjectively, as social cognitive biases may play a role in determining who is identified as a "top performer," the profiling approach can then serve to perpetuate bias in the organization.

Future Directions

This study is not without limitations. It analyzed data from a single concurrent validation study with a limited sample size. There are additional profiling and machine learning approaches that have not been evaluated in this study; more research is needed to explore additional approaches and replicate these findings in additional datasets, across context and assessment types, with larger sample sizes enabling an empirical analysis of the potential bias created by the various approaches. Future studies should also include additional methods for selecting top performers and extend the methodology to include Monte Carlo simulations.

Conclusion

In conclusion, results of this study show that the traditional approach to validation generally results in stronger observed estimates of predictive validity and greater differences in performance ratings for those who are classified as top performers by their managers compared to those who are not, suggesting that the traditional approach is most effective. These findings, combined with the other important considerations outlined above, lend strong support for traditional validation as a more rigorous and defensible method than profiling overall. These findings also shed light on the opportunity that new machine learning techniques present as a promising alternative, but they should be evaluated carefully before they are adopted.

REFERENCES

Alexander, L., Mulfinger, E., & Oswald, F. L. (2020). Using big data and machine learning in personality measurement: Opportunities and challenges. European Journal of Personality. doi: 10.1002/per.2305

Borra S. & Di Ciaccio A. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Computational Statistics and Data Analysis, 54(12), 2976–2989. doi: 10.1016/j.csda.2010.03.004

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. Journal of Applied Psychology, 100(2), 431–449. doi: 10.1037/a0038047

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. doi: 10.1023/a:1010933404324

Chang, C. & Lin, C. (2018). LIBSVM: A library for support vector machines. Retrieved from: http://www.csie.ntu.edu.tw/~cjlin/libsvm

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Erlbaum.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor & Department of Justice. (1978). Uniform guidelines on employee selection procedures. Retrieved from http://uniformguidelines.com/uniform-guidelines.html#:~:text=Uniform%20Guidelines%20on%20Employee%20Selection%20Procedures%201%20Section,Use%20of%20other%20validity%20studies.%20More%20items...%20

Fernandez-Araoz, C. (2020, May). Now is an unprecedented opportunity to hire great talent. Harvard Business Review. Retrieved from: https://hbr.org/2020/05/now-is-an-unprecedented-opportunity-to-hire-great-talent

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. Biometrika, 10 (4), 507–521. doi: 10.2307/2331838

Friedman, J., Hastie, T., & Tibshirani, R., (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1), 1-22. doi: 10.18637/jss.v033.i01

Gartner. (2018). Diversity and inclusion build high-performance teams. Retrieved from: https://www.gartner.com/smarterwithgartner/diversity-and-inclusion-build-high-performance-teams/

Gatewood, R. D., Feild, H. S., & Barrick, M. R. (2008). Human resource selection (6th edition). Thomson/South-Western.

Gonzalez, M. F., Capman, J. F., Oswald, F. L., Theys, E. R., & Tomczak, D. L. (2019). "Where's the I-O?" artificial intelligence and machine learning in talent management systems. Personnel and Assessment Decisions, 5(3), 33–44. doi: 10.25035/pad.2019.03.005

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. Journal of Applied Psychology, 75(3), 334. doi: 10.1037/0021-9010.75.3.334

Johnson, J. W., Steel, P., Scherbaum, C. A., Hoffman, C. C., Jeanneret, P. R., & Foster, J. (2010). Validation is like motor oil: Synthetic is better. Industrial and Organizational Psychology: Perspectives on Science and Practice, 3(3), 305–328. doi:10.1111/j.1754-9434.2010.01245.x

Kelley T. L. (1939). The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 30(1), 17–24. doi: 10.1037/h0057123

Kuhn, M., and Johnson, K. (2013). Applied predictive modeling. Springer New York. doi:10.1007/978-1-4614-6849-3.

Kulas, J. T. (2013). Personality-based profile matching in personnel selection: Estimates of method prevalence and criterion-related validity. Applied Psychology: An International Review, 62(3), 519–542. doi: 10.1111/j.1464-0597.2012.00491.x

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. Psychological Methods, 7(1), 19. doi: 10.1037/1082-989x.7.1.19

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill, Inc.

Paulhus, D. L., Robins, R. W., Trzesniewski, K. H. & Tracy, J. L. (2004). Two replicable suppressor situations in personality research. Multivariate Behavioral Research, 39, 303–328. doi: 10.1207/s15327906mbr3902_7

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. Organizational Research Methods, 21(3), 689–732. doi: 10.1177/1094428117697041

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. Journal of Applied Psychology, 104(10), 1207–1225. doi: 10.1037/apl0000405

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. Journal of Applied Psychology, 61(4), 473–485. doi: 10.1037/0021-9010.61.4.473

Schölkopf, B., Smola A., Williamson, R., & Bartlett, P. L. (2000). New support vector algorithms. Neural Computation, 12, 1207–1245. doi: 10.1162/089976600300015565

Society for Industrial and Organizational Psychology, Inc. (2018). Principles for the validation and use of personnel selection procedures (5th Ed.). SIOP. doi: 10.1017/iop.2018.195

Society for Industrial and Organizational Psychology. (2020). Top 10 workplace trends for 2020. Retrieved from: https://siop.inloop.com/en/article/14903/top-10-workplace-trends-for

Sullivan, J. (2014). The power has shifted to the candidate, so current recruiting practices will stop working. Retrieved from: https://www.ere.net/the-power-has-shifted-to-the-candidate-so-current-recruiting-practices-will-stop-working/

Talent Board. (2019). 2018 Talent Board North American candidate experience research report. Retrieved from: https://www.thetalentboard.org/wp-content/uploads/2019/02/2018_Talent-Board-NA-CandE-Research-Report_FINAL_2619.pdf

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. Personnel Psychology, 61(4), 871–925. doi: 10.1111/j.1744-6570.2008.00133.x

RECEIVED 02/14/20 ACCEPTED 11/11/20