Personnel Assessment and Decisions

Volume 8 | Issue 1 Article 7

2022

Comparing Empirically Keyed and Random Forest Scoring Models in Biodata Assessments

Mathijs Affourtit

Kristin S. Allen
SHL

Craig M. Reddock

Paul M. Fursman
SHL

Recommended Citation

Affourtit, Mathijs; Allen, Kristin S.; Reddock, Craig M.; and Fursman, Paul M. (2022) "Comparing Empirically Keyed and Random Forest Scoring Models in Biodata Assessments," *Personnel Assessment and Decisions*: Number 8 : Iss. 1 , Article 7.
DOI: https://doi.org/10.25035/pad.2022.01.007
Available at: https://scholarworks.bgsu.edu/pad/vol8/iss1/7

Comparing Empirically Keyed and Random Forest Scoring Models in Biodata Assessments

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

1. SHL

KEYWORDS

machine learning, validation, random forest, personnel selection

ABSTRACT

Effective pre-hire assessments impact organizational outcomes. Recent developments in machine learning provide an opportunity for practitioners to improve upon existing scoring methods. This study compares the effectiveness of an empirically keyed scoring model with a machine learning, random forest model approach in a biodata assessment. Data was collected across two organizations. The data from the first sample (*N*=1,410), was used to train the model using sample sizes of 100, 300, 500, and 1,000 cases, whereas data from the second organization (*N*=524) was used as an external benchmark only. When using a random forest model, predictive validity rose from 0.382 to 0.412 in the first organization, while a smaller increase was seen in the second organization. It was concluded that predictive validity of biodata measures can be improved using a random forest modeling approach. Additional considerations and suggestions for future research are discussed.

Properly developed and validated prehire selection systems can have an impact on important organizational outcomes, such as increased individual performance (Schmidt & Hunter, 1998), organizational performance (Lievens et al., 2020), and increased diversity (Sackett & Roth, 1996), to name just a few. Recent trends in technology are contributing to a number of positive developments in talent assessment, such as a more flexible and accessible assessment process (Mead et al., 2014), and the possibility of a more positive candidate experience (Miles & McCamey, 2018; Pulakos & Kantrowitz, 2016). Practitioners are taking advantage of advances in machine learning and analytics to improve the science that underlies prehire assessments such as biodata in an effort to maximize predictive validity (Pulakos & Kantrowitz, 2016).

In recent years machine learning approaches have been gaining popularity in the field of testing and assessment (e.g., Gonzalez et al., 2019). Putka et al. (2018) highlighted that the advances made in machine learning could be of value for I-O psychology and described methods in non-mathematical language. Recent research has found that machine learning techniques can even outperform traditional and profiling methods when used as a strategy for demonstrating the predictive criterion-related validity of assessments (Allen et al., 2020; Putka et al., 2018; Putka & Oswald, 2016).

When new assessments are developed, the response options are often empirically keyed (Cucina et al., 2012) to maximize predictive validity. Empirically keying responses for an item involves assigning a weight to each response option based on the correlation of each response option with a criterion of interest (i.e., job performance). Tests scored using this empirical keying methodology have been found to be more predictive of performance than tests scored with rationally keyed responses (e.g., Cucina et al., 2012; Devlin et al., 1992).

Moving beyond applications of empirically keyed scoring in assessment development, machine learning models can be developed and trained on assessment data to predict job-related criteria, such as manager ratings of

Corresponding author:
Mathijs Affourtit
Author Email: mathijsaffourtit@gmail.com

job performance, given sufficient data. These methods use candidate responses to assessment items as features and fit complex models to predict job-related criteria, picking up on patterns in the data that otherwise may not be detected when examining the predictive validities of single response options alone. Cross-validation techniques, such as k-fold cross-validation and Monte Carlo cross-validation, can be used to evaluate the effectiveness of the algorithm. In cross-validation, an algorithm is trained on a subset of data and applied to a different subset of the same dataset to evaluate how well the algorithm works (Hastie et al., 2009).

A machine learning model well-suited to handle a range of response options is the *random forest model* (Breiman, 2001). A random forest predictor fits multiple decision trees to the training data, with a limited number of features being available at each split to increase the diversity of each tree. Each tree is trained on a bootstrapped sub sample. The predictions of these individual trees are combined into an overall prediction (Breinman, 2001; Kuhn & Johnson, 2013; Putka et al., 2018). The random forest model is a particularly advantageous modeling technique for this supervised learning task using multiple choice, categorical responses, as it is highly versatile and can handle a large number of features (Strobl et al., 2009). Furthermore, it is robust to outliers and works with nonlinear data (Breinman, 2001; James et al., 2017; Kuhn & Johnson, 2013). Machine learning models are typically less interpretable than empirical scoring methods due to their black box nature. However, it is possible to derive a measure of feature importance from a random forest model (Hastie et al., 2009). Feature importance describes the relevance of an individual predictor to predicting the outcome variable, similar to how a regression coefficient represents the strength of the relationship between a predictor and the dependent variable in regression analysis. Although not used in this study, feature importance could be used to select the most predictive items when developing a measure by prioritizing the inclusion of the most predictive items.

Present Study

This paper seeks to extend the findings by Putka et al. (2018) highlighting the potential value of machine learning methods for I-O psychology, by investigating the effectiveness of applying machine learning scoring algorithms to assessment development efforts and the practical implications of doing so. This study evaluates whether a machine learning random forest scoring model can outperform an empirically keyed scoring approach in terms of predictive validity on a biodata assessment in an applied setting. Biodata assessments are known to strongly and consistently predict job performance (Becton et al., 2009; Bliesener, 1996; Breaugh et al., 2014; Rothstein et al., 1990; Schmidt & Hunter, 1998) and have shown incremental validity when used with other testing content (Allworth & Hesketh, 2000). Additionally, biodata assessments have been found to predict a number of important work outcomes (Hunter & Hunter, 1984; Reilley & Chao, 1982), including turnover and organizational commitment (Becton et al., 2009), and are relevant across job roles and levels (Schmitt et al., 1984). Furthermore, Breaugh et al. (2014) found the use of biodata in personnel selection resulted in minimal subgroup score differences for both gender and race/ethnic groups, whereas Bradburn and Schmitt (2019) demonstrated that pairing biodata with a cognitive ability test can reduce its negative impact on selection ratios for some minority groups. Given its various qualities, biodata has proven to be a powerful and useful tool for I-O psychologists. For this reason, the present study focuses on a popular and well-established biodata measure (SHL, 2015).

When sufficient data are available, implementing scoring keys based on random forest models may improve validity. However, the amount of data required to execute such modeling effectively is highly dependent on the specific circumstances. For example, Allen et al. (2020) found that when using a random forest model, results depended greatly on the sampling method used. Because the need for large amounts of data is a practical limitation of using machine learning techniques, this study will also investigate, in an applied manner, how much data would be required to achieve comparable prediction using empirical keyed scoring approach and random forest machine learning scoring models. This research question is intended to help I-O psychology practitioners better understand the circumstances under which the use of machine learning techniques may be a viable alternative to traditional methods.

To summarize, this study has the following objectives: first, to evaluate the effectiveness of machine learning models to score biodata assessment; second, to assess whether these approaches could be used as an alternative to empirical keyed scoring; finally, to review the impact of sample size on validity of the models in an effort to provide guidance on the sample size required to reach equivalent or improved validity.

METHOD

Sample

Data were available through research collaborations with two organizations in the healthcare and insurance industries. The data were cleaned and removed if managers were unfamiliar with the employee's behavior and work performance. More specifically, incumbent data were removed if managers responded "Cannot Rate" when asked how familiar they are with the employee's job performance. Participants were also removed if they were missing testing data.

After cleaning, the data contained a total of 1,934 cases, with 1,410 cases (1,566 before cleaning) from the first organization and 524 cases (574 before cleaning) from the second organization. The models were trained using data from the first organization with the data from the second organization serving as an external benchmark. The data contained item-level responses, manager ratings, and limited demographic data. Of the total sample, 35% (685) were male, 37% (721) were female, and 27% (528) preferred not to answer when asked to indicate their gender. When asked to indicate their race, 27% (528) participants preferred not to answer, whereas others identified as White (47%; 928), Black/African American (10%; 203), Hispanic/Latino (7.5%; 145), Asian 3.7%; (72), two or more races (2.4%; 47), Native Hawaiian/Other Pacific Islander (0.3%; 7), and American Indian/Alaska Native (0.2%; 4). Information about age was available for 1,406 respondents, with the average age being 42.4 years (*SD*: 10.5).

Measures

The biodata assessment used in this study was designed to predict professional success across a wide variety of jobs and industries. Assessment items ask candidates to answer questions related to their past achievements, social orientation, work style, strengths, and work aspirations (SHL, 2015). The items were written to measure job-related factors found to be important across multiple job analyses. Using data from thousands of employees across multiple organizations, responses were empirically keyed based on the degree to which they were statistically related to job performance ratings. The 15 most predictive items were selected from a larger item bank to comprise the final operational version of the assessment. This methodology resulted in a multidimensional construct model that is driven by both research on a theoretical construct model of widely applicable job-related behaviors and their empirical relationships with performance outcomes. The items ask candidates to complete the sentence in the item stem by selecting a response option to report the frequency, quality, or experience related to their past performance on a variety of job-related behaviors. Each item consists of four to six response options (mode = 5, mean = 5.3). The final version of the assessment takes approximately 4 minutes to complete.

The assessment has been found to have adequate test–retest reliability with a correlation of .68 between multiple administrations. The assessment has been found to demonstrate construct validity across multiple studies, exhibiting positive meta-analytic correlations with tests measuring similar constructs including sales potential (r = .22), safety judgment (r = .22), management judgment (r = .22), learning potential (r = .21), and management potential (r = .19), and negligible relationships with measures of unrelated constructs such as contact center skills including data entry accuracy (r = .05), data entry speed (r = .04), working with information (r = .04), tactful problem solving (r = .03), and navigation (r = .02).

Performance Measure

Data were consolidated from eight concurrent validation studies spanning two client partners. Performance was measured using a variety of managerial job performance ratings on a 5-point scale including four items in the first organization and six items in the second organization assessing overall performance (e.g., “If you had your choice of job candidates, would you hire this employee again?”). Responses to these items were combined to form a global performance composite. All four overall job performance ratings included in the first clients global composite score overlap with four of the six items included in the second clients composite. The two additional items rated the overall match between each associate’s ability and job requirements as well as the overall match between each associate’s values and the organization’s culture. For the purpose of this study, we will consider the data collected from the two clients as two separate samples.

Study Design

The data from the first client were used as the main sample, whereas the data from the second client were used as an external benchmark. Using a Monte Carlo cross-validation design with 100 repetitions, the main sample was split into a training and a test sample. The training sample was used to train the models, whereas the test sample was used to establish the performance of those models. The performance of the models was evaluated by computing the correlations between the scores provided by the model and the manager ratings. Correlation was chosen as the evaluation criteria as it is generally the most common measure in I-O psychology to demonstrate evidence of criterion-related validity (Gatewood et al., 2008).

Up to 1,000 candidates were selected from the main sample for training, leaving 410 candidates in the test sample. The models were built using different subsets of the training sample of increasingly smaller size to determine the effect of the training sample size on model performance. In order to avoid introducing variance in the validity of the models due changes in the test sample composition, the holdout sample was kept constant while reducing the training sample. Specifically, the data from candidates who were removed from the training sample went unused rather than being added to the test sample. Four models were created using the following sample sizes: 1,000, 500, 300, and 100. The baseline validity was calculated using the existing empirically keyed scoring. This process was repeated 100 times using different training cases.

Before training the models, item responses were converted to dummy variables where each of the dummy variables indicated whether the candidate had selected a specific response within the item. The random forest models were trained on the dummy variables. The random forest model was implemented using the randomForest package in R (R Core Team, 2017). Due to the relatively small data size (compared to typical machine learning applications), the number of trees for each model was reduced to 150 (from the default of 500). Other tuning parameters were left at their default values.

For each repetition, the data from the second organization was used as an external benchmark or additional holdout sample. Using this as a separate sample, rather than merging the two samples, allows us to evaluate the model's performance using an independent sample, which gives further insight in the generalizability of the model. This external benchmark was not split into training and test samples, as all training was performed using the main sample. After running all repetitions, the results were aggregated, and mean, standard deviation, and minimum and maximum validity were calculated for each method and sample.

RESULTS

The aggregated performance metric for the empirically keyed and random forest scoring models can be found in Table 1 and displayed in Figures 1 and 2. Using the existing scoring method, the average validity in the main sample was 0.382. The average validity of the random forest models in the main sample ranged from 0.355 to 0.412, with the highest average validity showing for the model that used the largest training sample and the validity decreasing as the training sample got smaller. Training a model using 300 cases led to a comparable validity (0.394) as the empirical keyed scoring method.

When looking at the model performance in the external sample, the same general trend holds, but the validities are generally lower compared to the main sample. The average validity of the existing scoring method is 0.205 with the random forest validities ranging from 0.174 to 0.218. For the external sample, training a model using 500 cases led to a comparable validity (0.205) as the existing scoring method. Note that because the composition of the test external sample does not change, the empirical validity was the same across repetitions.

DISCUSSION

This paper set out to, in an applied setting, (a) evaluate the effectiveness of machine learning models, specifically random forest models, in terms of predictive validity on psychometric assessments; (b) assess whether these scoring methods can be used as an alternative to an empirically keyed scoring approach; and (c) review the impact of sample size on the validity in an effort to establish guidance regarding the sample size required to achieve equivalent or greater validity. The study used a within-sample cross-validation approach as well as an external sample collected from a different organization to validate the results.

There are many considerations to take into account

TABLE 1.
Correlations With Global Performance Ratings (100 Repetitions)

Sample	Method	Training *N*	*Mean*	*SD*	*Min*	*Max*
Main	Empirically keyed scoring	n/a	0.382	0.034	0.278	0.444
	Random forest	1000	0.412	0.036	0.285	0.517
		500	0.406	0.037	0.295	0.504
		300	0.394	0.036	0.281	0.494
		100	0.355	0.053	0.203	0.493
External	Empirically keyed scoring	n/a	0.205	0	0.205	0.205
	Random forest	1000	0.218	0.017	0.18	0.253
		500	0.205	0.025	0.141	0.264
		300	0.195	0.029	0.13	0.264
		100	0.174	0.041	0.082	0.262

FIGURE 1.

Distribution of empirical and RF model validities in the main sample

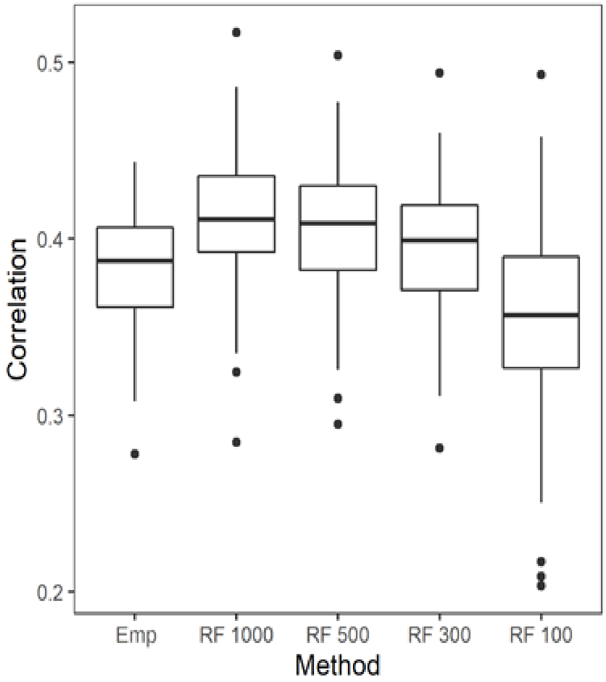
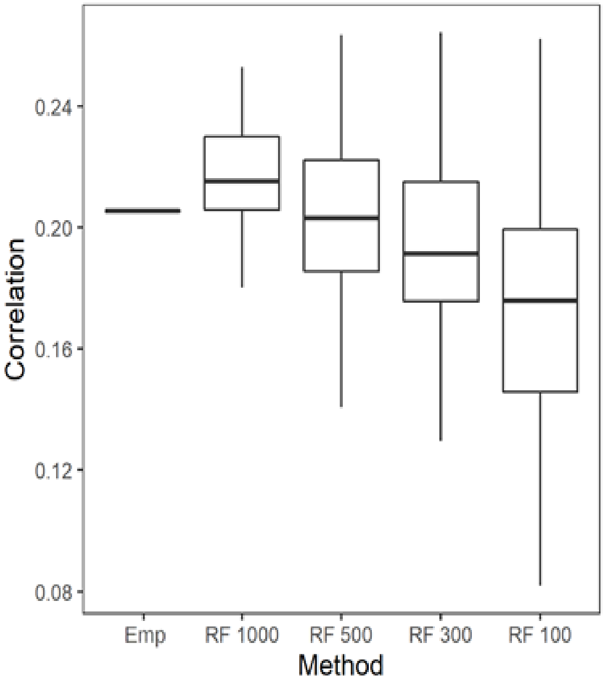


FIGURE 2.

Distribution of empirical and RF model validities in the external benchmark sample



when developing scoring models for psychometric assessments, such as validity, sample size required to develop and validate the assessment, as well as the explainability and interpretability of the approach. The results of the study demonstrated that random forest techniques can produce scoring models that outperform empirically keyed scoring when using within-sample cross-validation and achieve equivalent validity when using a relatively small sample of 300–500 cases. However, when using an external benchmark sample collected in a different organization, a much larger sample was needed to approximate equivalence. When using the largest sample available in this study, the random forest model average validity was larger than the empirically keyed scoring model, with validities ranging from 0.180 to 0.235.

Different scoring methodologies have different levels of explainability, with empirically keyed scoring being the most interpretable and explainable, as a candidate's response can directly be associated with an impact on their overall score. Another commonly used scoring model is item response theory (IRT). The increased complexity of IRT leads to the scoring method being less interpretable and explainable. The impact of a candidate's response on the overall score is less transparent, as their responses on one item are evaluated in relation to their responses on the other items. This is even more pertinent when items are presented using a computer adaptive approach. However, IRT parameters, such as difficulty and discrimination, associated with each item give some indication how responses affect the final score.

Similarly to item response theory models, when using a random forest scoring model, a candidate's response on an item can no longer be associated with an impact of their overall score, due to the increased complexity of the models involved and the resulting nonlinearity between item responses and the overall score. However, random forest models provide an estimation of the importance of each of the questions in terms of impact on the overall score, which gives insight into which items are most impactful on the overall score.

Limitations

This study focused on estimating the feasibility of using machine learning to create and improve scoring models for biodata assessments. Through the use of a large cross-validation sample, this study was able to review the impact of the size of the training sample on the validity of the assessment. However, this study is far from encompassing all aspects that are relevant to the question of the value of utilizing machine learning in psychometric assessment. This study used the random forest model, which is highly versatile, can effectively handle non-normal data, and is somewhat robust to overfitting to the data. There are, however, many more machine learning models in existence that could be applicable options for developing psychometric measures. Future studies should aim to evaluate a broader range of empirical keying approaches, such as those outlined by Cucina and colleagues (2012), as well as additional machine learning models. Future research should focus on those with transparent and interpretable scoring methods, as explainable AI is likely to be an important factor in years to come (Arrieta et al., 2020). Furthermore this study focused on the use of empirical keyed scoring methods and did not evaluate other established scoring models such as IRT and multidimensional IRT models (Brown & Maydeu-Olivares, 2011; Hambleton et al., 1991). Future studies should aim to include a broader comparison to existing and established scoring methodologies. Additionally, this study used an existing measure with a fixed set of items; items were not selected as part of this study. Future studies could make use of feature importance measures provided by random forest models to select items. Finally, this study used a biodata measure; future studies should evaluate a broader range of assessment types to determine whether the findings reported in this paper are generalizable beyond biodata assessments.

Another ongoing challenge for I-O practitioners conducting validation studies are sample sizes. Though this study relied on a relatively large sample of almost 2,000 cases combined across both organizations, the validity of random forest models could likely be increased further by using even larger samples. Additionally, the models were trained using a single sample collected from one organization. Using samples from a more diverse set of organizations to train the model would likely increase the generalizability. Given the ongoing nature of working online and increasing data awareness of organizations across the world, we can expect that using larger samples will be easier to obtain over time (Parkins, 2017). Conversely, it is important to learn more about the lower bound of sample sizes for the efficacy of random forest models, as well as other machine learning techniques. Organizations would benefit greatly from knowing more about what a "too small" sample might be for scoring assessments and what factors might mitigate or exacerbate the accuracy of the models (e.g., personality vs biodata, job level). Future studies should continue to look at a range of sample sizes as well as other factors to help provide additional guidance regarding best practices.

Ultimately, when developing assessments to predict job performance, machine learning techniques can be one more option for practitioners to consider. Given the promising results of this study and the growing enthusiasm around machine learning, practitioners need to be informed about appropriate practical applications of machine learning in the context of personnel selection.

REFERENCES

Allen, K. S., Affourtit, M., & Reddock, C. 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, 6(3), Article 2. https://doi.org/10.25035/pad.2020.03.002

Allworth, J., & Hesketh, B. (2000). Job requirements biodata as a predictor of performance in customer service roles. International Journal of Selection and Assessment, 8(3), 137–147. https://doi.org/10.1111/1468-2389.00142

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Becton, J. B., Matthews, M. C., Hartley, D. L., & Whitaker, D. H. (2009). Using biodata to predict turnover, organizational commitment, and job performance in healthcare. International Journal of Selection and Assessment, 17(2), 189–202. https://doi.org/10.1017/S1833367200000638

Bliesener, T. (1996). Methodological moderators in validating biographical data in personnel selection. Journal of Occupational and Organizational Psychology, 69, 107-120. https://doi.org/10.1111/j.2044-8325.1996.tb00603.x

Bradburn J., & Schmitt N. (2019). Combining cognitive and noncognitive predictors and impact on selected individual demographics: An illustration [J]. International Journal of Selection and Assessment, 27(1): 21-30. https://doi.org/10.1111/ijsa.12234

Breaugh, J., Labrador, J., Frye, K., Lee, D., Lammers, V., & Cox, J. (2014). The value of biodata for selecting employees: Comparable results for job incumbent and job applicant samples? Journal of Organizational Psychology, 14(1), 40-51.

Breiman, L. (2001). Random forests. Machine Learning 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. Educational and Psychological Measurement, 71(3), 460-502. https://doi.org/10.1177/0013164410375112

Cucina, J. M., Caputo, P. M., Thibodeaux, H. F., & Maclane, C. N. (2012). Unlocking the key to biodata scoring: A comparison of empirical, rational, and hybrid approaches at different sample sizes. Personnel Psychology, 65, 385. https://doi.org/10.1111/j.1744-6570.2012.01244.x

Devlin, S. E., Abrahams, N. M., & Edwards, J. E. (1992). Empirical keying of biographical data: Cross-validity as a function of scaling procedure and sample size. Military Psychology, 4, 119-136. https://doi.org/10.1207/s15327876mp0403_1

Gatewood, R. D., Feild, H. S., & Barrick, M. R. (2008). Human resource selection. 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 Assessment and Decisions, 5(3), 5. https://doi.org/10.25035/pad.2019.03.005

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). Fundamentals of item response theory. Sage Publications, Inc.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Prediction, inference and data mining. Springer Verlag.

Hunter, J.E., & Hunter, R.F. (1984). Validity and utility of alternate predictors of job performance. Psychological Bulletin, 96, 72–98. https://doi.org/10.1037/0033-2909.96.1.72

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning with applications in R (7th edition). Springer.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.

Lievens, F., Sackett, P. R., & Zhang, C. (2020). Personnel selection: A longstanding story of impact at the individual, firm, and societal level. European Journal of Work and Organizational Psychology, 1-12. https://doi.org/10.1080/1359432X.2020.1849386

Mead, A. D., Olson-Buchanan, J. B., & Drasgow, F. (2014). Technology-based selection. In M. D. Coovert & L. F. Thompson (Eds.), The psychology of workplace technology (p. 21–42). SIOP Organizational Frontiers Series. Routledge/Taylor & Francis Group.

Miles, S. J., & McCamey, R. (2018). The candidate experience: Is it damaging your employer brand? Business Horizons, 61(5), 755-764. https://doi.org/10.1016/j.bushor.2018.05.007

Parkins, D. (2017). The world's most valuable resource is no longer oil, but data. Economist, 6. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Pulakos, E. D., & Kantrowitz, T. (2016). Choosing effective talent assessments to strengthen your organization. SHRM Foundation's Effective Practice Guidelines Series. https://www.shrm.org/foundation/news/documents/choosing%20effective%20talent%20assessments.pdf

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. https://doi.org/10.1177/1094428117697041

Putka, D. J., & Oswald, F. L. (2016). Implications of the big data movement for the advancement of I-O science and practice. In S. Tonidandel, E. B. King, & J. M. Cortina (Eds.), Big data at work: The data science revolution and organizational psychology (pp. 181–212). Routledge/Taylor & Francis Group.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Reilly, R. R., & Chao, G. R. (1982). Validity and fairness of some alternative employee selection procedures. Personnel Psychology, 35, 1-62. https://doi.org/10.1111/j.1744-6570.1982.tb02184.x

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? Journal of Applied Psychology, 75(2), 175–184. https://doi.org/10.1037/0021-9010.75.2.175

Sackett, P. R., & Roth, L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. Personnel Psychology, 49(3), 549-572. https://doi.org/10.1111/j.1744-6570.1996.tb01584.x

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin, 124(2), 262. https://doi.org/10.1037/0033-2909.124.2.262

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37(3), 407-422. https://doi.org/10.1111/j.1744-6570.1984.tb00519.x

SHL (2015). Professional Potential Technical Manual. Thames Ditton, UK: SHL.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. Psychological Methods, 14, 323–348. https://doi.org/10.1037/a0016973

RECEIVED 05/06/2021 ACCEPTED 01/31/2022