CHRO ASSOCIATION

SIOP Foundation
BRIDGING SCIENCE AND PRACTICE

How to Evaluate Artificial Intelligence (AI) Based Employment Tools from Vendors

Organizations often do not have the expertise to adequately vet AI-based tools; instead, they trust vendors and simply benchmark with what others are doing.

This guide aims to support the evaluation of AI-based employment products from vendors.

Over the last decade, many start-ups have emerged, claiming to sell AI-based tools that solve employment challenges and often dismissing over a century of Industrial and Organizational (I/O) psychologists' research and practice in employment testing.

The novelty of these products and vendor confidence often far outweigh their organizational effectiveness, the time-consuming start-up costs to deploying technology, and legal defensibility. At the same time, organizations often lack the expertise to properly vet these tools and instead rely on vendor claims or peer benchmarking.

In this context, we:

- Outline where AI may add value in employment settings.
- Provide a checklist of key questions, risks, and requirements to help leaders evaluate vendor claims and make sound, legally defensible investments.
- Include a glossary and research references for readers who want additional background.

Print the **checklist.**

Automating the analysis and scoring of text data is probably the biggest opportunity of AI in Human Resources (HR) to date.

Why AI May Be Especially Advantageous When Making Employment Decisions

Maximizes Efficiency in High-Volume Hiring

One of the greatest payoffs from AI lies in high-volume hiring or employee management scenarios. Processing large candidate pools can increase efficiencies and rapidly offset AI tool investments, particularly in labor-intensive tasks like prescreening resumes, scoring applications, and automating initial interviews.

Untapped HR Data

AI excels at extracting, analyzing, and scoring underutilized text data from candidate submissions and employee records. The use of additional, job relevant information improves the prediction of future performance. Organizations generate vast, rich text supplies during recruitment and throughout the employee life cycle, yet human evaluation is often biased, time-intensive, and costly—leaving critical insights underused or ignored.

Smarter Predictive Hiring

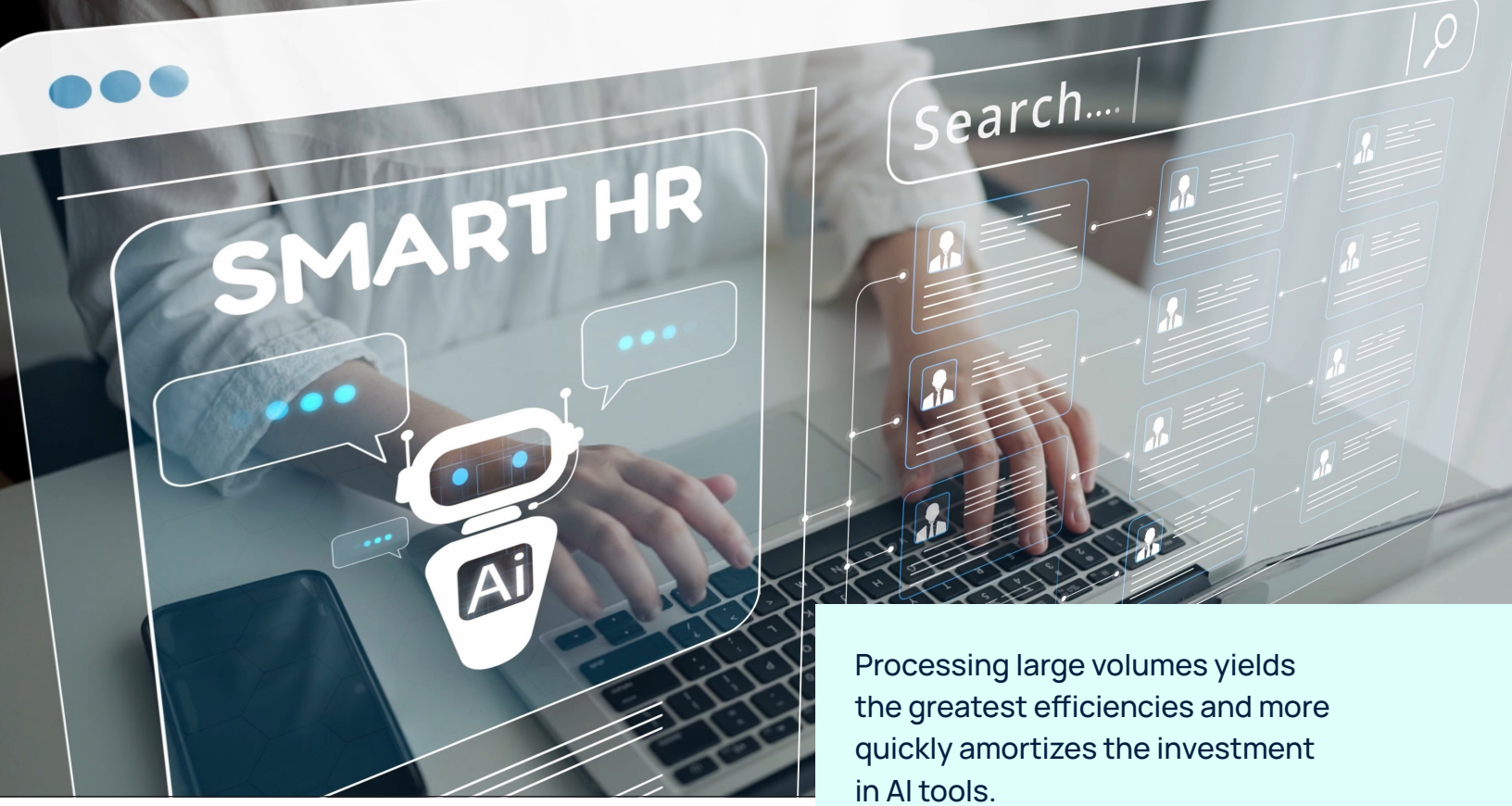
Natural Language Processing (NLP) enables a more standardized, quantifiable analysis of this data to inform hiring. When AI technologies are selected and implemented appropriately, they can enhance predictive validity (job performance forecasting in I/O psychology) over traditional methods by incorporating novel text-based signals and skills, leveraging greater data volumes, and applying more optimal statistical approaches.

Consistent AI Scoring

Deterministic AI models ensure identical scores for the same inputs and can help mitigate bias if well designed. Conversely, model trained on skewed, irrelevant data require monitoring for algorithmic bias.

Fair, Lawful AI Use

Well-designed deterministic AI models avoid amplifying passing rate gaps or adverse impact across gender/racial subgroups beyond data baselines. Research shows they can reduce disparities, boost validity, and assess evenly distributed skills—without offsetting real differences in job-relevant knowledge, abilities, or credentials. Vendors promising AI-driven reductions via score adjustments warrant suspicion, as this is unlawful.



Processing large volumes yields the greatest efficiencies and more quickly amortizes the investment in AI tools.

Uses of AI in Employment with Research Support

Research suggests there are many ways that AI can offer value to employment decisions. Below are examples that have been successful. Note, however, that not all vendors have developed a successful product.

Scoring employment applications and resumes.

To date, this is one of the most common uses. Many studies have supported AI's effectiveness for this purpose. Specifically, deterministic AI models can score the information as well as or better than a human rater.

However, accurate scoring of applications and resumes may be limited by the quality and thoroughness of the information collected on the application regardless of whether AI or a human conducts the evaluation.

Scoring automated interviews.

Similar to application forms and resumes, AI technologies are found to score as accurately or better than a human interviewer and much faster.

The algorithm must be trained on a set of interview questions. The accuracy of generic vendor products not specifically trained for the organization should be statistically evaluated, not assumed.

Regardless of the choice of interview questions, consistency in the interview questions used across candidates and over time is required to be most effective.

Automating the data capturing and scoring for other types of assessments.

Examples include collecting text-based answers like responses to questionnaires, and quantitative answers.

“Gamifying” assessments to make them more appealing.

Sometimes AI is used to “gamify” assessments presumably to make them more enjoyable for candidates, but there is not yet sufficient evidence that gamified assessments are consistently more effective than traditional forms of assessment and in some cases, they may be less so compared to traditional forms of assessment. Users of gamified assessments should carefully evaluate their effectiveness in the context of their jobs.

These uses of AI may be unique to the specific organization and are probably limited to high-volume selection contexts for development to be cost effective.

Creating test and interview questions.

Probabilistic AI models (e.g., Large Language Models or LLMs) can develop questions that may be indistinguishable from those developed by subject matter experts (SMEs), or these models can assist SMEs in generating initial assessment content. The use of AI to create test and interview content can save the high labor and time costs required historically. The quality of such questions should be confirmed by final SME review, pilot testing, and psychometric analyses.

Analyzing jobs and job requirements.

Probabilistic AI models can facilitate job analyses to make this important task much easier. However, do not rely solely on such tools. Use them to provide supplementary information as part of the job analysis, and do not assume the accuracy of the information produced without SME review.

AI can help identify applications and resumes that match the job requirements, including those specified in the job description.

The primary sources of information should be local job analyses and archival data such as O*NET that are verified locally.

Assisting with recruiting.

AI can help identify applications and resumes that match the job requirements, including those specified in the job description, which can save recruiter time compared to manual review.

Some AI tools can even search the internet for candidates with posted resumes that match the required skills and encourage them to apply. Additionally, these tools can recommend different jobs within the organization that may be a better fit or have a greater need for applicants who do apply, based on their applications or resumes.

There may be a risk of overreliance on such tools because they are often integrated into the Applicant Tracking Systems (ATS) and automatically presented to the recruiter. The impact of AI outreach tools on different applicant groups should be monitored and, where needed, supplemented with lawful, job-related outreach to broaden the qualified applicant pool while complying with all nondiscrimination and merit-based requirements. In addition, they should be validated, ideally predicting job performance but minimally predicting recruiter judgments of resumes and applications.

We offer a list of questions organizational leaders should ask vendors trying to sell AI-based products.

Questions to Ask Vendors and Recommended Requirements for Using AI

Considering what AI-based tools can offer, what the potential advantages to organizations are, and what research suggests AI does well, we offer a list of questions organizational leaders should ask vendors trying to sell AI-based products.

These questions are mostly focused on employee selection, but many can also be applied when AI is being considered for some other aspect of human resource management.

1 What selection-relevant job applicant skills or other attributes are measured by the AI-based assessment?

Vendor responses should include documented empirical evidence (not just words that sound good), which may be based on how it was developed or on research studies they have conducted.

Review the content of the test yourself and have testing experts and SMEs do so. Ensure that what the tool measures aligns with the requirements of the jobs, which should be based on a job analysis. What the AI-based assessment measures and how it works should be explained in a way that is understandable to a layperson.

Further, AI models that are largely "black box" (meaning that its content cannot be directly observed) should have explicit validation evidence that is generalized to your organizational context. Those that mainly depend on the assertions of the vendor should be **avoided.**

The vendor team should have experts in testing, measurement, psychometrics, employee selection, and related topics.

2 What research and process were used to motivate and develop the AI-based assessment?

What was the central framework behind the assessment? What steps were followed in its development?

Aside from using AI, the steps normally include those in any other measurement development situation: defining the attribute to be measured, reviewing the relevant research literature, developing assessment items, administering them to a sample of candidates to analyze and refine the assessment, determining scoring and administrative details, collecting criterion data such as job performance to validate the assessment, relating the job requirements to the assessment, and documenting all steps and results in a report.

Especially ensure that the sample used to develop the assessment is comparable to employees or applicants in your organization and not based on convenience samples such as Prolific or MTurk, which are commerically available sources of data collected from unknown respondents for a fee, because these are often low-quality and participants do not represent the actual context (e.g., hiring, promoting).

This background information on the test is important because it provides information on the logic, proper steps, thoroughness, and rigor with which the test was developed, and ultimately its value and defensibility.

If you need assistance in evaluating a test, contact an expert who is well-versed in test development, including AI models, and validation. Remember too that data from other companies can be useful in making a decision about the choice of the AI-based tool; however, other organization's data will not be sufficient in the event of a legal challenge.

3 What education and experience does the vendor have that qualifies them to develop and sell this assessment?

The vendor team should have experts in testing, measurement, psychometrics, employee selection, and related topics. The educational background and work experience of the people who developed the test are extremely important because they reflect domain expertise and practical experience in selection and HR, which s typically not possessed by information technology vendors.

To have confidence in the assessment and in the event of a legal challenge, you want developers who have education and experience related to assessment development and validation.

4 What evidence is there of the reliability of the AI-based assessment?

Reliability refers to the consistency of assessment results and the accuracy of the scores. For example, would candidates achieve roughly the same scores if they were assessed on a different day? Would alternative versions of the assessment yield similar scores, as might be needed if candidates retest or reveal the content to other candidates? Does the assessment measure a single job-relevant skill or attribute or many?

The reliability of results is a vital prerequisite for validity, fairness, and legal defensibility of the AI-based product as it is with any selection procedure.

5 What evidence is there for the validity of this AI-based assessment?

Validity refers to the accuracy of the predictions made from a test score. There are several ways to demonstrate validity, but in the hiring context with AI, validity usually refers to *criterion-related validity,* which is how well the scores statistically predict future job performance, turnover, or other outcome (criterion) of importance to the organization.

The vendor should be able to provide many forms of validity evidence, such as multiple studies and multiple criteria in organizations similar to yours to suggest that support for criterion-related validity is realistic in your organization as well.

Insist on documented empirical research-based validity, not vague reports of success stories or testimonials from past clients, which often dominate marketing materials. Importantly, do not accept exaggerated promises; even if results are provided, ensure those results are not cherry-picked. Vendors may only present the most successful studies, not the ones that failed.

Importantly, require a validity study in your own organization. Validity evidence from other organizations can be informative, but it should not be solely relied on. In fact, a vendor's claims of validity are an insufficient defense by themselves when an organizational hiring process shows adverse impact (i.e., large differences in passing rates by subgroup), according to the *Uniform Guidelines on Employee Selection Procedures,* which is a set of standards from the federal government usually critically important to a legal defense.

Another source of validity evidence is content validity evidence that evaluates the link between the content of the test and the requirements of the job. However, content validity evidence of an AI-based selection tool by itself is probably not sufficient without criterion-related validity evidence.

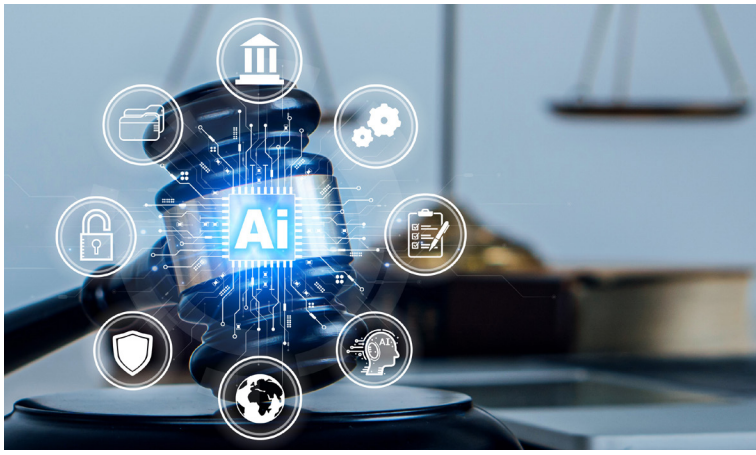
AI-based selection scores should not be used to automatically screen candidates without human oversight.

6 What research has been conducted on gender and race differences on the AI-based assessment scores?

Gender and race differences in average scores are not illegal if the assessment is job-related (i.e., the assessment is valid). Conversely, the lack of mean differences on an AI-based assessment does not necessarily mean it is a good assessment because it still might not be reliable or valid.

Subgroup differences in average scores are not uncommon in valid assessments. An employer should know what subgroup differences exist in scores from AI-based assessments in advance to anticipate potential legal exposure and challenges to achieving a representative workplace. Evidence from other organizations provided by the vendor will not be a sufficient defense if challenged.

Finally, AI-based selection scores should not be used to automatically screen candidates without human oversight. There should always be a "human in the loop" to ensure the process is running properly and to perhaps consider other information in the decision.



7 What is the business case and return on investment?

Adoption of AI-based assessments should consider at least four categories of costs: (a) the direct cost of purchasing the product, which may include license fees, per-use charges, set-up costs; (b) the investment by the organization in implementing the product, which may include costs associated with information technology (IT) staff time to integrate the product into the organization's IT infrastructure, HR staff and management time to integrate the product into the hiring process and coordinate the project, time and expense for training employees to use the process, and expert time (e.g., selection scientists, lawyers); (c) costs for conducting job analysis and validation research; and (d) costs of administering the process over time, such as maintenance (by vendor and IT staff), HR time addressing issues raised by hiring staff and candidates (e.g., use of scores, policies, potential errors), especially initially, any special reporting and monitoring time.

Depending on the instrument, the benefits may include improvements to the quality of new hires, reduced staff time in implementing assessments, faster selection decision making, or increased access and fairness.

8 What documentation exists on the AI-based assessment?

Appropriate documentation will normally include technical reports on the development of the assessment, the job analysis that identified the KSAOs, the validation of the assessment in the client organization, and any special research conducted by the vendor.

These reports should be accessed, read, understood, and retained. People with a background in employment testing can help you determine the quality and adequacy of the research supporting the use of the assessment.

9 Are there accommodations for candidates with disabilities?

As with any other selection procedures, simple accommodations such as font sizes, color contrast, logical layouts and instructions, and a method for requesting additional assessment time or adjustments for other needs should be considered.

An organization may also want to consider alternative assessments for candidates unable to complete the procedure even with these accommodations.

10 Are data privacy and security requirements met in record keeping?

As with other HR data, the vendor must ensure that candidate and employee data they collect as part of the employment process meet data privacy laws that are applicable in the location where they are collected. The vendor should protect data by using firewalls to keep the data secure from hacking, collecting only necessary information, keeping data for only the necessary period of time, using data only for intended purposes, omitting personally identifying information in research activities outside the employment decision, and allowing access to data only by those with a legitimate need to know.

Finally, determine who owns the data collected and why. Is it the property of the organization or the vendor? Who owns the data after the contract ends?

Additional Questions:

11 **What recommendations or experiences does the vendor have with managing stakeholder reactions proactively?**

A wide range of people, including HR staff, hiring managers, and candidates, may experience initial skepticism of AI-based assessments due to a lack of understanding. Ask the vendor how other organizations have addressed common questions that arise.

12 **Will the vendor monitor legislation for new laws that create requirements for users of AI in employment?**

Federal, state, and local laws pertaining to the use of AI assessments in employment settings are evolving, and more will likely emerge in the future. In addition, these laws may vary and sometimes conflict. What is permissible in one location may be prohibited in another. Vendors and users of their products must comply with relevant laws. Like previous technological advancements, it is likely that public reaction will become more positive as we continue to understand AI's value in the employment space.

13 **Will the vendor monitor the growing candidate use of Large Language Models (LLMs), such as ChatGPT, to generate answers to questions and provide other textual employment information?**

Vendors should monitor LLM use with detection software and evaluate the consequences of LLM use, such as whether candidates that use LLMs are selected more often. The use of LLMs is especially a concern with unproctored administration of automated interviews, as well as in all applications and interviews. Although no vendor can guarantee it captures all instances of AI-generated responses, it is important to monitor because the rate of usage is growing.

The next few pages contain a **checklist** of these questions, plus red flags and requirements, to use as a tool for evaluating vendor claims.

Checklist

Click on the check boxes and type notes directly in the document.

CHECKLIST OF QUESTIONS TO EVALUATE VENDORS WHO PROVIDE ARTIFICIAL INTELLIGENCE-BASED SELECTION PROCEDURES

QUESTION	RED FLAGS	REQUIREMENTS	NOTES
☐ 1 **What selection-relevant job applicant skills or other attributes are measured by the AI-based assessment?**	• Attributes not highly job related, thus not benefiting the organization or legally defensible if challenged. • "Black box" models that cannot be understood or evidence is not shared.	• Empirical evidence based on research from development process and in the organization. • Base on job analysis. • Review of assessment content by company.	
☐ 2 **What research and process were used to motivate and develop the AI-based assessment?**	• Illogical or lacking a scientific central framework. • Development process vague, difficult to understand, or not logically relevant to jobs.	• Central framework is logical. • Proper assessment development steps followed. • Used sample(s) similar to the job to be staffed.	
☐ 3 **What education and experience does the vendor have that qualifies them to develop and sell this assessment?**	• Developed by entrepreneurs or information technology specialists without domain knowledge.	• Education and experience related to assessment development and validation.	
☐ 4 **What evidence is there of the reliability of the AI-based assessment?**	• Scores are inconsistent, indicating substantial error of measurement.	• Research-based statistical evidence of reliability.	
☐ 5 **What evidence is there for the validity of this AI-based assessment?**	• Exaggerated claims. • Based on success stories or testimonials. • Cherry-picking results. • Evidence is only based on other organizations.	• Statistical data showing prediction of future job performance, turnover, or other criteria, especially in your organization. • Multiple studies and criteria.	

Checklist *(continued)*

CHECKLIST OF QUESTIONS TO EVALUATE VENDORS WHO PROVIDE ARTIFICIAL INTELLIGENCE-BASED SELECTION PROCEDURES

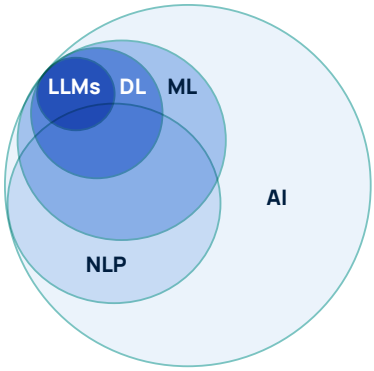
	QUESTION	RED FLAGS	REQUIREMENTS	NOTES
☐	**6** **What research has been conducted on gender and race differences on the AI-based assessment scores?**	• Unanticipated or unexplainable diversity differences. • Lack of diversity differences due to lack of validity. • Evidence entirely based on other organizations.	• Data on percentage of highly skilled candidates by subgroups. • Diversity differences based on true differences in skills between subgroups. • Human in the loop.	
☐	**7** **What is the business case and return on investment?**	• Unanticipated or excessive costs. • Vague or assumed benefits.	• Clearly known costs in advance, including hidden costs like implementation and maintenance. • Clear evidence/data on benefits, both financial and non-financial.	
☐	**8** **What documentation exists on the AI-based assessment?**	• Lack of documentation. • Only documents are presentations or marketing materials.	• Technical reports describing development and validation. • Reports on any research in organization. • Documents read and understood.	
☐	**9** **Are there accommdations for candidates with disabilities?**	• Lack of accommodations or needs not unanticipated.	• Accommodations incorporated into design. • Easy to implement (e.g., extra time).	
☐	**10** **Are data privacy and security requirements met in record keeping?**	• Privacy or security concerns or exposures.	• Privacy laws met. • Adequate security. • Proper ownership of data.	

Checklist *(continued)*

CHECKLIST OF QUESTIONS TO EVALUATE VENDORS WHO PROVIDE ARTIFICIAL INTELLIGENCE-BASED SELECTION PROCEDURES

	QUESTION	RED FLAGS	REQUIREMENTS	NOTES
	ADDITIONAL QUESTIONS			
☐	**11** **What recommendations or experiences does the vendor have with managing stakeholder reactionsproactively?**	• Product likely to require extensive management of stakeholder reactions. • Vendor has little useful insight.	• Product unlikely to provoke negative stakeholder reactions. • Useful suggestions for managing reactions.	
☐	**12** **Will the vendor monitor legislation for any new laws that create new requirements for users of AI in employment?**	• Vendor lacks deep knowledge of laws. • No formal plan for monitoring new laws.	• Vendor knowledgeable and experienced in meeting laws. • Vendor monitors legal developments and informs customers.	
☐	**13** **Will the vendor monitor for the growing use of Large Language Models (LLMs) like ChatGPT by candidates to generate answers to questions and provide other textual employment information?**	• Vendor product susceptible to AI-generated answers or other forms of cheating and lacks good action plan for addressing.	• Vendor product prevents or statistically checks for AI-generated text and has plan for addressing.	

Glossary of Technical Terms



RELATIONSHIPS AMONG TERMS

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
NLP	Natural Language Processing
LLMs	Large Language Models

Note: Adapted from K. Silva (2011)

Artificial Intelligence (AI)
Methods that simulate intelligent human behavior, such as learning, decision making, sensing, and predicting.

Black Box
Colloquialism to describe how the mathematical operations of some ML algorithms cannot be examined and understood directly.

Deep Learning (DL)
A subset of ML; uses neural networks and transformers that have many internal layers of analysis.

- **Transformers**
  A type of deep-learning architecture that uses advanced attentional mechanisms to improve understanding of human language, even if words are far apart in a sentence.

Deterministic versus Probabilistic AI Models
Deterministic models produce the exact same scores or other output each time, given the same inputs; probabilistic models like Large Language Models produce similar but not identical output each time, given the same inputs.

Features
Variables extracted and scored by a machine learning model; in NLP, this includes n-grams.

Ground Truth
Criterion used to train machine learning models to predict, such as past hiring decisions.

Large Language Models (LLMs)
Language-based models that generate text data; trained in millions or billions of parameters.

Machine Learning (ML)
A primary AI method including algorithms that refers to sophisticated methods for detecting patterns and can 'learn' from these patterns and new data by adjusting the model to improve its effectiveness.

- **Supervised Machine Learning**
  An ML model trained to detect the relationships between variables and a criterion (ground truth), and then replicate those relationships on new, unseen data.

- **Unsupervised Machine Learning**
  An ML model the inductively uncovers patterns in the data; a human makes sense of the patterns; there is no criterion in unsupervised ML

Natural Language Processing (NLP)
Algorithms specifically designed to enable machines to understand and generate human language.

- **Bag-of-Words (BoW)**
  An NLP vectorization method that represents text data as unordered words (e.g., "woman leads people" is the same as "leads woman people").

- **n-grams**
  A historical baseline method of NLP where terms are extracted from a corpus of text; can be single terms ("leader") or multi-terms (e.g., bi-grams, "strong leader")

- **Word Embeddings**
  An NLP vectorization method that represents text data while considering syntax, unlike BoW.

Other Resources

The citations to research articles and related documents below provide further information on using AI in employee selection from the scientific literature. To facilitate finding articles related to your interest, the categories are based on applications of AI to types of selection procedures or related topics. The list of citations are relatively complete with regard to employee selection journals in I/O psychology, but there may be relevant studies in other related disciplines (e.g., computer science, data science, information processing, statistics, linguistics, cognitive science). Moreover, new studies are coming out with great frequency, so the findings are evolving.

SCORING APPLICATIONS

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101(7),* 958-975.

Campion, E. D., Campion, M. A., Johnson, J., Carretta, T. R., Romay, S., Dirr, B., Deregla, A., & Mouton, A. (2024). Using natural language processing to increase prediction and reduce subgroup differences in personnel selection decisions. *Journal of Applied Psychology, 109(3),* 307-338.

Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., Koohifar, F., Yankov, G., ... & Newton, C. (2023). Improving measurement and prediction in personnel selection through the application of machine learning. *Personnel Psychology, 76(4),* 1061-1123.

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. J*ournal of Applied Psychology, 104(10),* 1207-1225.

SCORING INTERVIEWS

Liff, J., Mondragon, N., Gardner, C., Hartwell, C. J., & Bradshaw, A. (2024). Psychometric properties of automated video interview competency assessments. *Journal of Applied Psychology, 109(6),* 921-948.

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology, 107(8),* 1323-1351.

Stevenor, B. A., Hickman, L., Zickar, M. J., Wimbush, F., & Beck, W. (2024). Validity evidence for personality scores from algorithms trained on low-stakes verbal data and applied to high-stakes interviews. I*nternational Journal of Selection and Assessment, 32(4),* 544-560.

SCORING ASSESSMENTS

A) Multiple-Choice Tests

Hickman, L., Dunlop, P. D., & Wolf, J. L. (2024). The performance of large language models on quantitative and verbal ability tests: Initial evidence and implications for unproctored high-stakes testing. *International Journal of Selection and Assessment, 32(4),* 499-511.

Landers, R. N., Auer, E. M., Dunk, L., Langer, M., & Tran, K. N. (2023). A simulation of the impacts of machine learning to combine psychometric employee selection system predictors on performance prediction, adverse impact, and number of dropped predictors. *Personnel Psychology, 76(4),* 1037-1060.

al-Qallawi, S., & Raghavan, M. (2022). A review of online reactions to game-based assessment mobile applications. *International Journal of Selection and Assessment, 30(1),* 14-26.

B) Constructed Response Assessments

Hickman, L., Herde, C. N., Lievens, F., & Tay, L. (2023). Automatic scoring of speeded interpersonal assessment center exercises via machine learning: Initial psychometric evidence and practical guidelines. I*nternational Journal of Selection and Assessment, 31(2),* 225-239.

Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., Koohifar, F., Yankov, G., ... & Newton, C. (2023). Improving measurement and prediction in personnel selection through the application of machine learning. *Personnel Psychology, 76(4),* 1061-1123.

Thompson, I., Koenig, N., Mracek, D. L., & Tonidandel, S. (2023). Deep learning in employee selection: Evaluation of algorithms to automate the scoring of open-ended assessments. *Journal of Business and Psychology, 38(3),* 509-527.

C) Personality Assessments

Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., ... & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. J*ournal of Applied Psychology, 108(8),* 1277-1299.

Hernandez, I., & Nie, W. (2023). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology, 76(4),* 1011-1035.

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology, 107(8),* 1323-1351.

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2023). A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing. *Journal of Business and Psychology, 38(1),* 163-190.

Song, Q., Oh, I. S., Kim, Y., & So, C. (2024). Revisiting the nature and strength of the personality–job performance relations: New insights from interpretable machine learning. *Journal of Applied Psychology.*

Stevenor, B. A., Hickman, L., Zickar, M. J., Wimbush, F., & Beck, W. (2024). Validity evidence for personality scores from algorithms trained on low-stakes verbal data and applied to high-stakes interviews. *International Journal of Selection and Assessment, 32(4),* 544-560.

D) Job Performance Assessments

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology, 71(3),* 299-333.

Speer, A. B. (2021). Scoring dimension-level job performance from narrative comments: Validity and generalizability when using natural language processing. *Organizational Research Methods, 24(3)*, 572-594.

Speer, A. B., Schwendeman, M. G., Reich, C. C., Tenbrink, A. P., & Siver, S. R. (2019). Investigating the construct validity of performance comments: Creation of the great eight narrative dictionary. *Journal of Business and Psychology, 34,* 747-767.

ASSISTING JOB ANALYSIS

Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., Koohifar, F., Yankov, G., ... & Newton, C. (2023). Improving measurement and prediction in personnel selection through the application of machine learning. *Personnel Psychology, 76(4),* 1061-1123.

Putka, D. J., Oswald, F. L., Landers, R. N., Beatty, A. S., McCloy, R. A., & Yu, M. C. (2023). Evaluating a natural language processing approach to estimating KSA and interest job analysis ratings. *Journal of Business and Psychology, 38(2),* 385-410.

ASSISTING OTHER HR PURPOSES

Banks, G. C., Woznyj, H. M., Wesslen, R. S., Frear, K. A., Berka, G., Heggestad, E. D., & Gordon, H. L. (2019). Strategic recruitment across borders: An investigation of multinational enterprises. *Journal of Management, 45(2),* 476-509.

Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., & Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through AI capability framework. *Human Resource Management Review, 33(1),* 100899.

Green, J. P., Dalal, R. S., Fyffe, S., Zaccaro, S. J., Putka, D. J., & Wallace, D. M. (2023). An empirical taxonomy of leadership situations: Development, validation, and implications for the science and practice of leadership. *Journal of Applied Psychology, 108(9),* 1515-1539.

Guo, F., Gallagher, C. M., Sun, T., Tavoosi, S., & Min, H. (2024). Smarter people analytics with organizational text data: Demonstrations using classic and advanced NLP models. *Human Resource Management Journal, 34(1),* 39-54.

Kotlyar, I., & Krasman, J. (2022). Virtual simulation: New method for assessing teamwork skills. *International Journal of Selection and Assessment, 30(3),* 344-360.

Kumar, L. S., & Burns, G. N. (2024). Determinants of safety outcomes in organizations: Exploring O* NET data to predict occupational accident rates. *Personnel Psychology, 77(2),* 555-594.

Min, H., Peng, Y., Shoss, M., & Yang, B. (2021). Using machine learning to investigate the public's emotional responses to work from home during the COVID-19 pandemic. Journal of Applied Psychology, 106(2), 214-229.

Min, H., Yang, B., Allen, D. G., Grandey, A. A., & Liu, M. (2024). Wisdom from the crowd: Can recommender systems predict employee turnover and its destinations? Personnel Psychology, 77(2), 475-496.

Song, Q. C., Shin, H. J., Tang, C., Hanna, A., & Behrend, T. (2024). Investigating machine learning's capacity to enhance the prediction of career choices. *Personnel Psychology, 77(2),* 295-319.

Speer, A. B., Perrotta, J., Tenbrink, A. P., Wegmeyer, L. J., Delacruz, A. Y., & Bowker, J. (2024). Turning words into numbers: Assessing work attitudes using natural language processing. *Journal of Applied Psychology, 108(6),* 1027-1045.

Zhang, C., Yu, M. C., & Marin, S. (2021). Exploring public sentiment on enforced remote work during COVID-19. *Journal of Applied Psychology, 106(6),* 797-810.

FAIRNESS, BIAS, AND DIVERSITY

Campion, E. D., Campion, M. A., Johnson, J., Carretta, T. R., Romay, S., Dirr, B., Deregla, A., & Mouton, A. (2024). Using natural language processing to increase prediction and reduce subgroup differences in personnel selection decisions. *Journal of Applied Psychology, 109(3),* 307-338.

Chekili, A., & Hernandez, I. (2024). Demographic inference in the digital age: Using neural networks to assess gender and ethnicity at scale. *Organizational Research Methods, 27(2),* 301-328.

Hickman, L., Huynh, C., Gass, J., Booth, B., Kuruzovich, J., & Tay, L. (2024). Whither bias goes, I will go: An integrative, systematic review of algorithmic bias mitigation. J*ournal of Applied Psychology, 110(7),* 979-1000.

Rottman, C., Gardner, C., Liff, J., Mondragon, N., & Zuloaga, L. (2023). New strategies for addressing the diversity–validity dilemma with big data. J*ournal of Applied Psychology, 108(9),* 1425-1444.

Wang, W., Dinh, J. V., Jones, K. S., Upadhyay, S., & Yang, J. (2023). Corporate diversity statements and employees' online DEI ratings: An unsupervised machine-learning text-mining analysis. *Journal of Business and Psychology, 38(1),* 45-61.

Zhang, N., Wang, M., Xu, H., Koenig, N., Hickman, L., Kuruzovich, J., ... & Kim, Y. (2023). Reducing subgroup differences in personnel selection through the application of machine learning. *Personnel Psychology, 76(4),* 1125-1159.

GUIDELINES

Society for Industrial and Organizational Psychology (SIOP). (2023). *Considerations and Recommendations for the Validation and Use of AI-Based Assessments for Employee Selection.* Bowling Green: Author.

Society for Industrial and Organizational Psychology (SIOP). (undated). SIOP Employment Testing Guide. Adapted from www.SIOP.org.

REVIEWS OF RESEARCH AND COMMENTARIES

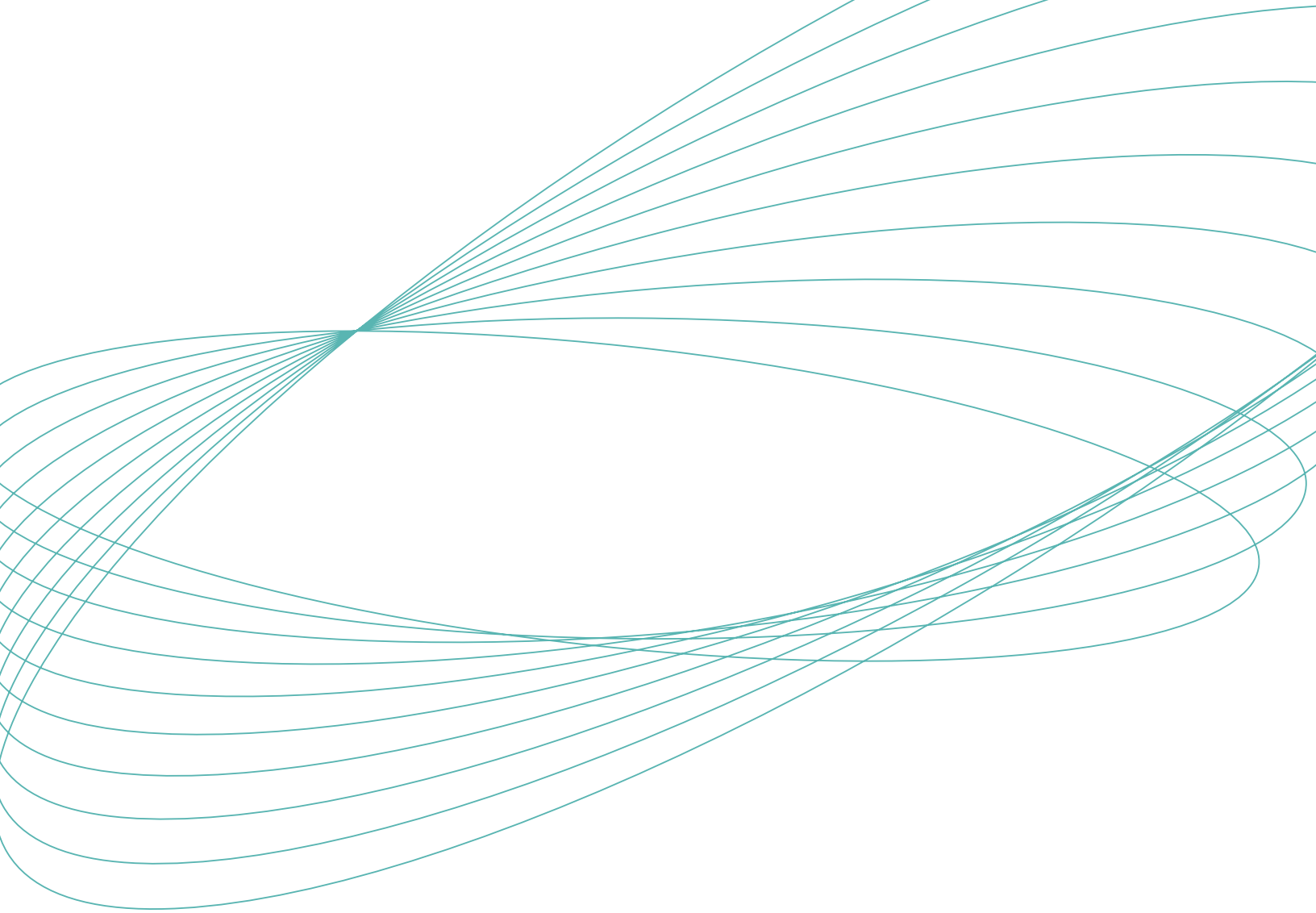
Campion, E. D., & Campion, M. A. (2020). Using computer-assisted text analysis (CATA) to inform employment decisions: Approaches, software, and findings. *Research in Personnel and Human Resources Management, 38,* 285-325.

Campion, M. A., & Campion, E. D. (2023). Machine learning applications to personnel selection: Current illustrations, lessons learned, and future research. *Personnel Psychology, 76(4),* 993-1009.

Campion, E. D., & Campion, M. A. (2024). Impact of machine learning on personnel selection. *Organizational Dynamics, 53(1),* 101035.

Campion, E. D., & Campion, M. A. (2025). A review of text analysis in human resource management research: Methodological diversity, constructs identified, and validation best practices. Human Resource Management Review, 35 (2), 101078.

Campion, M. A. (2026). Can legal and professional personnel selection principles be met with machine learning (artificial intelligence)? *Human Resource Management, 65,* 235–255.



This paper is the result of a partnership between the CHRO Association and the SIOP Foundation and is the product of a panel of experts formed by the SIOP Foundation.

CHRO Association is the only independent community dedicated to Chief Human Resource Officers (CHROs). A trusted network of peers, experts and advocates building the world's future-ready workforces and a powerful force for growth, CHRO Association supports those leaders that boards and CEOs rely on to build organizations that thrive through change. We shape HR policy, regulation, and workforce strategies to create stronger businesses, people and society.

The goal of the working group was to summarize what scientific support for the use of AI in employment exists and provide a checklist to use with vendors when considering one of these tools. Mike Campion and Emily Campion were the primary authors of this paper. They were supported by other members of the working group, including Mirian Graddick-Weir, Neil Morelli, Cole Napper, Fred Oswald, Rob Ployhart, David Rodriguez, and Nancy Tippins.

The SIOP Foundation is the philanthropic arm of the Society for Industrial and Organizational Psychology (SIOP) and actively partners with other organizations like the CHRO Association to promote the development and application of the science of industrial and organizational psychology in the workplace. In addition to these partnerships, the SIOP Foundation maintains an active program of awards, scholarships, and research grants. More information about the SIOP Foundation can be found at its website: https://www.siop.org/foundation/. Inquiries regarding the SIOP Foundation's work can be directed to SIOPFoundation@siop.org.