From Intuition to Algorithm
Why AI is Becoming Indispensable in Personnel Selection

Marcus Heidbrink, Florian Feltes Zortify S.A., Luxembourg Published: 2025, zortify.com

Abstract

Despite decades of criticism, methods such as unstructured interviews, CV analysis and, in particular, assessment centres continue to dominate the field of personnel selection, even though they often fail to meet key psychometric requirements and offer little predictive value despite the high level of effort involved. In view of the growing importance of evidence-based, fair and scalable selection processes, the use of artificial intelligence (AI) is increasingly coming into focus. This article first highlights the shortcomings of traditional selection procedures based on current meta-analyses. It then systematically presents the scientific evidence on language-based AI procedures in personality diagnostics. Using the *Zortify* procedure as an example, results on objectivity, reliability and validity are presented. On this basis, it becomes clear that AI-based diagnostics not only complement traditional methods, but can also surpass them in key aspects such as objectivity, long-term measurement stability, predictive validity and resource-efficient scalability. Finally, requirements for a scientifically sound, ethically acceptable and data protection-compliant implementation are discussed. The article is intended as a plea for evidence-based, future-proof personnel selection. Given the proven potential, foregoing AI-based methods is not only backward-looking, but increasingly irresponsible.

Keywords: Personnel selection, artificial intelligence, assessment centre, psychometric criteria, language-based personality assessment

Key findings:

- Traditional selection procedures often do not meet minimum psychometric standards.
- AI-supported procedures offer diagnostic added value when they are evidence-based.
- The Zortify assessment procedure meets key psychometric requirements.
- Processes such as Zortify offer the opportunity to make personnel selection more fair, accurate and, at the same time, more cost-effective.

Heidbrink, M., & Feltes, F. (2025). *From Intuition to Algorithm: Why AI is Becoming Indispensable in Personnel Selection* [White paper]. Zortify. https://zortify.com

1. Status Quo of Personnel Selection: Limitations of Traditional Methods

In an age of demographic change, intensified competition for skilled workers and increasing regulation such as data protection regulations and the European AI Act, organisations are under more pressure than ever to make well-founded, efficient, legally compliant and fair personnel decisions. Accordingly, the call for objective, evidence-based selection procedures is growing louder. Nevertheless, traditional methods such as unstructured interviews, the pure analysis of application documents or assessment centres continue to dominate corporate practice. However, a growing number of scientific studies indicate that these methods have significant weaknesses, particularly with regard to psychometric quality criteria such as objectivity, reliability and validity. As a result, they are largely responsible for wrong decisions.

Three recent reviews provide guidance. First, the comprehensive international review by Van Iddekinge, Lievens and Sackett (2023), which takes a critical look at traditional and new personnel selection methods with a focus on validity, fairness, candidate experience and technological developments. Secondly, the study by Armoneit, Schuler and Hell (2020), which focuses on the German market and reports a long-term survey to show how the use of selection methods has developed. Thirdly, the meta-analysis by Sackett et al. (2022), which summarises specific validity coefficients for different selection procedures.

The three studies reveal a central paradox: methods frequently used in personnel selection have only limited psychometric quality, while more innovative, evidence-based approaches remain largely unused. The methods themselves are developing only slowly. The potential of modern diagnostic methods, such as AI-supported analyses, is not yet being exploited, although it could contribute substantially to improving fairness, accuracy and predictive power.

1.1 Commonly Used Selection Methods

Despite the well-known shortcomings in fairness, validity, reliability and objectivity, traditional procedures dominate selection practices. The study by Armoneit et al. (2020) shows that German companies continue to rely primarily on the following methods:

1. analysis of application documents (82% of companies),

2. unstructured interviews (34%), structured interviews (73%),
3. work samples (46%)
4. and assessment centres (38%).
5. Test-based procedures such as personality tests (19%; online 24%)
6. or performance tests (24%; online 16%)

have played a minor, albeit emerging, role to date. Van Iddekinge et al. (2023) report a similar ranking and low level of innovation in selection methodology based on international data.

1.2 Psychometric Quality of Traditional Selection Methods

The scientific rigor of a personnel assessment method lies at its core: its objectivity, reliability and validity. The analyses by Armoneit et al. (2020), Sackett et al. (2022) and Van Iddekinge et al. (2023) show that it is those procedures that are most widely used that often do not meet the minimum psychometric requirements. Assessment centres play a special role among traditional procedures: despite numerous empirical indications of their limited validity, high costs and methodological fallibility, they are considered the diagnostic "gold standard" in many organisations.

Application documents are used in almost every selection process. Nevertheless, their validity is well below the acceptable range. This is indicated by correlations with job success of r = .07 (for professional experience in years) to r = .22, in the best case of systematically evaluated CV data (Sackett et al., 2022). The assessment is usually intuitive, without uniform criteria, and subject to considerable observation errors. Reliability is correspondingly low. The appeal of this method lies primarily in its practicality, not in its diagnostic value.

Interviews, especially when conducted in an unstructured manner, have low objectivity and only moderate validity (r = .19; Sackett et al., 2022). Only structured interviews achieve slightly higher values (r = .42; Sackett et al., 2022), however, are less commonly used in practice. The high level of subjectivity and personal influence of the interviewers, the wording of the questions and the lack of evaluation schemes remain central weaknesses of unstructured interviews.

Work samples offer moderate validity (r = .33; Sackett et al., 2022). They are particularly effective in roles that are closely related to practical tasks and are considered acceptable by applicants. However, their use often involves increased organisational effort.

Assessment centres are often considered the "gold standard" of personnel selection, although they have fallen short of this claim for years. The reviews cited indicate that they do not live up to this reputation empirically or practically (Armoneit et al., 2022, Van Iddekinge et al., 2023; Sackett et al., 2022). In many cases, the effort involved is disproportionate to the diagnostic value. Studies indicate that even highly structured assessment centres are only of limited use for making valid statements about professional success due to observation errors, low criterion validity (r = 0.29; Sackett et al., 2022) and high costs. They are highly dependent on the design, the training of the observers and the standardisation of the exercises (Gaugler et al., 1987). Observers regularly underestimate the distorting effects of their subjectivity. The actual informative value of these procedures is limited.

Personality tests based on self-reporting generally achieve acceptable levels of objectivity and reliability. Their predictive validity for professional success is also moderate (r = .25 - .30; Sackett et al., 2022). However, studies on work performance, counterproductive behaviour or teamwork (Barrick & Mount, 1991; Zell & Lesick, 2022) emphasise the relevance of personality traits for professional success. However, self-report personality tests are susceptible to socially desirable response behaviour, especially in selection settings (Birkeland et al., 2006; Kowalski et al., 2018), which can contribute to limited predictive validity. Van Iddekinge et al. (2023) therefore recommend combining these tests with observation-based tests to minimise bias and capture more authentic personality profiles, including through the use of validated AI-based methods.

1.3 Reasons for the Dominance of Weaker Methods

But why do some methods persist despite their empirically proven weaknesses? The answer lies in a combination of practicality, social acceptance and institutional barriers. Armoneit et al. (2020) emphasise that unstructured interviews and CV analyses are considered particularly simple, cost-effective and flexible, and are also culturally accepted; HR professionals find these methods intuitive and interactive, which increases their acceptance by both companies and applicants. Assessment centres are often considered particularly advantageous, despite being among the most complex and costly selection procedures and offering only moderate validity.

Van Iddekinge et al. (2023) add that many decision-makers overestimate the informative value of subjective procedures and avoid evidence-based procedures due to their complexity, perceived coldness or legal uncertainty. A report by Rahe & Rahe (2017) also shows that only 39% of the companies surveyed cite "scientific rigour"

as a key criterion for diagnostic procedures, while aspects such as comprehensibility (68%) and practicality (77%) are mentioned far more frequently.

These findings clearly show that the weaknesses of traditional selection procedures are well known. Despite their high costs, logistical complexity, and moderate validity, assessment centers and other traditional procedures remain in use, driven more by psychological convenience and institutional inertia than by psychometric merit.

1.4 Prospects for AI-Based Diagnostics

Against this backdrop, AI-supported processes are attracting growing attention in both research and practice. Van Iddekinge et al. (2023) see potential in the use of AI for personnel selection. By integrating empirically validated AI approaches, diagnostic procedures can achieve greater objectivity, reliability, and validity, ultimately minimizing personnel selection errors. Language data in particular offer a new way to access latent characteristics such as personality, motivation or cognitive style, which is less susceptible to conscious distortion (e.g. Moreno et al., 2021; Yarkoni, 2010).

The potential benefits of using AI in personnel selection include:

1. the reduction of subjective bias in assessments,
2. access to characteristics that are difficult to measure,
3. the analysis of unstructured data,
4. processing large amounts of data, and
5. scalability and cost efficiency.

The aspects of validity mentioned above (1 and 2) and the analytical possibilities of modern data processing (3 and 4) consistently lead to the argument of scalability and cost efficiency (5). Traditional methods such as assessment centres incur high costs per candidate and, due to their susceptibility to error, do not protect against expensive selection errors. In contrast, digitised, evidence-based AI processes enable location-independent, fully automated implementation with minimal resource use and a higher success rate in filling positions, particularly due to points 1 and 2, which increase predictive power. Particularly in terms of scalability and long-term cost reduction, evidence-based AI processes therefore offer a sustainable solution that can be both economically and diagnostically superior to traditional selection methods such as assessment centres.

However, for successful implementation, AI-supported processes must also undergo rigorous scientific validation processes and meet the same psychometric standards

as traditional instruments. In this sense, they can not only compensate for the existing weaknesses of traditional methods, but also usher in a new era of personnel selection, moving away from intuition and towards data-driven, fair and scalable decision-making.

2. Psychometric Quality of AI-Based Methods: A Review of Current Research

The findings outlined above demonstrate that widely used selection methods, including unstructured interviews, CV analyses or traditional assessment centres, tend to have low validity and are vulnerable to subjective biases. Assessment centres in particular have been shown to be particularly susceptible to observer bias, group effects and contextual influences. As a result, objective comparability between candidates is significantly limited. Despite their widespread use, they have methodological weaknesses that have been shown to lead to systematic errors in decision-making.

Against this backdrop, attention is increasingly turning to new, technology-based approaches, in particular methods based on artificial intelligence (AI). Language-based AI models that analyse natural language (NLP models) from application documents, interviews or open-ended responses promise a paradigm shift in this area. They have the potential to enable standardised, transparent and scalable diagnostic processes with the aim of minimising human bias, capturing personality traits more reliably and minimising personnel decision errors.

The following section provides a systematic overview of the empirical evidence on the use of AI-based methods in personnel selection, with a focus on validity, reliability, objectivity and practical applicability. It is based on current meta-analyses, systematic reviews and promising individual empirical studies from recent years.

2.1 Synthesis of Current Reviews on AI-Based Personality Diagnostics

Interest in AI-based methods for analysing personality has grown significantly. Numerous meta-analyses and systematic reviews have investigated the extent to which AI-supported methods can reliably capture personality traits, particularly on the basis of text data. The focus is usually on the Big Five personality dimensions, and less frequently on the Dark Triad.

The studies conducted to date paint a promising picture overall, even though the quality of the methods varies greatly. Since 2019, a large number of overview studies have been published, which are systematically summarised in this chapter. The analysed studies consist of two meta-analyses (Moreno et al., 2023; Koutsoumpis et al., 2022) and three reviews (Bhandarkar et al., 2024; Hashemi-Motlagh et al., 2025; Naz et al., 2025). They evaluate average validity, compare different model architectures and identify challenges such as a lack of standardisation and transparency. An overview of the analysed studies, their approach and key findings is provided in Table 1.

2.2 Psychometric Quality of AI-based Personality Diagnostics

According to these reviews, AI- and NLP-based methods are often effective in providing objective, reliable, and valid assessments of personality traits. Because they do not entirely rely on traditional self-reporting, these methods are considered promising in terms of ecological validity, standardisation and automation (Bhandarkar et al., 2024).

Validity. The empirical evidence to date on the validity of speech-based AI methods in personality diagnostics paints a promising but heterogeneous picture overall. Meta-analyses and systematic reviews consistently conclude that AI-based models can capture relevant aspects of personality, but that validity is moderate and highly context-dependent. The meta-analysis by Moreno et al. (2021) reports average predictive validity for Big Five traits in the range of $r \approx .26–.30$, with longer and more content-rich texts enabling better predictions. Koutsoumpis et al. (2022) supplement these findings with corrected effect sizes (ρ) between .08 and .14 (self-reports) and .18 to .39 (external report). The results underline that AI methods are capable of capturing relevant personality traits, even if they do not provide a comprehensive representation of traditional constructs.

Reliability. Initial data on reliability are also available: In large-scale studies with open text entries or AI-based chatbots, retest coefficients between $r = .48$ and $r = .70$ are reported, indicating moderate temporal stability (Park et al., 2015; Fan et al., 2023; Hickman et al., 2022). Hybrid methods that combine questionnaire data with NLP analyses achieve retest reliability between $r = .30$ and $r = .60$, depending on the personality trait in question (Koutsoumpis et al., 2024; Zhang et al., 2024).

Table 1

Overview of reviews on the quality of AI-based methods for measuring personality based on text

Authors	Method	Key findings
Moreno et al. (2021)	• Meta-analysis • 23 primary studies on AI-based personality diagnostics via text • Focus: Relationship between language data and the Big Five • Moderator analysis (e.g., text source, trait, model type) • Objective: determine predictive validity	• Average predictive validity: r ≈ .26–.30 for Big Five • Text source and trait moderate predictive validity (e.g., higher values for conscientiousness, lower for agreeableness) • Classic ML models (e.g., SVM, regression) show moderate performance • Validity improves with longer and more thematically broad text input
Koutsoumpis et al. (2022)	• Meta-analysis • 31 independent samples (~85,000 participants) • Focus: Correlations between AI-based categories and the Big Five • Distinction between self-description and external description • Effect sizes as corrected correlations (ρ) • Moderator analyses: platform, language, description type	• Methods show low to moderate correlations with Big Five (self-reports: | ρ| ≈ .08–.14; external reports: |ρ| ≈ .18–.39) • Higher validity for external reports than for self-reports • Context dependence is crucial: platform, language and communication goal influence results • Conclusion: AI captures aspects of personality, but does not provide a comprehensive representation
Hashemi-Motlagh et al. (2025)	• Review • Over 100 studies on text, audio and video analysis • Focus: Comparison of modalities and model architectures • Categorisation by input type, target feature, methodology • Objective: Overview of the state of the art and areas of application	• Multimodal methods (text + audio/video) show the highest validity • Transformer-based models consistently outperform classic models • Validation often inconsistent; recommendation for more standardised benchmarks
Naz et al. (2025)	• Review: methodological and technical overview • Focus: SVM, CNN, RNN, Transformer models • Comparison of methods, data types and feature engineering • Evaluation based on predictive performance, training data, evaluation metrics	• Transformer architectures (BERT, GPT) deliver above-average prediction quality • Traditional models (SVM, Random Forest) deliver inconsistent results • Call for context-sensitive, data protection-compliant applications in human resources
Bhandarkar et al. (2024)	• Review: Comparison of existing methods • Evaluation based on quality criteria (objectivity, validity, fairness, explainability) • Discussion of ethical challenges and bias risks	• Objectivity and automation as major strengths • Criticism of lack of explainability and psychological foundation • Plea for ethical minimum standards and psychological basis

Objectivity. AI-based personality assessment methods are considered objective in the literature because they are standardised, automated and free from human bias when the training data quality is high (Bhandarkar et al., 2024; Naz et al., 2025). Unlike traditional selection procedures such as unstructured interviews or assessment centres, where subjective assessments by observers can have a significant influence on the outcome, AI-supported systems are based on consistent algorithmic evaluations that are independent of the applicant or evaluator. AI-based methods use only text information that has been provided voluntarily. Since no photos, names or information on gender and age are required, this not only reduces susceptibility to implicit biases, but also ensures a high level of data protection and fairness. This form of standardisation increases comparability across individuals, contexts and time periods.

At the same time, several reviews point to necessary limitations: Actual objectivity depends largely on the quality and lack of bias in the training data and on the transparency of the models used (Hashemi-Motlagh et al., 2025; Bhandarkar et al., 2024). In particular, so-called black box models carry the risk that although the evaluation is automated, the underlying decision-making rules are not comprehensible. However, objectivity in the narrower psychometric sense requires that results are not only generated in a standardised manner, but are also intersubjectively comprehensible. The use of explainable AI is therefore increasingly being discussed as a key requirement.

2.3 Determinants of the Quality of AI-based Methods

The quality of AI-based personality diagnostics is influenced by a number of technical and psychometric factors:

- **Psychological foundation**: Methods based on validated personality models (e.g., Big Five) perform better consistently. Their theoretical clarity increases both validity and acceptance in application (Bhandarkar et al., 2024).
- **Text Length and Quality**: The length and semantic depth of the analysed texts are crucial. Longer, more personal texts with greater diversity of content enable more stable personality predictions (Moreno et al., 2021).
- **Linguistic and Cultural Robustness**: Many methods are primarily designed for English-language training data. Transferability to other languages and cultural contexts is often limited, which restricts international applicability (Koutsoumpis et al., 2022).

- **Model Architecture**: The most powerful predictor of quality is the model structure used. Transformer-based models such as BERT, RoBERTa or GPT show significantly better predictive performance than more traditional approaches (e.g. random forests, SVMs or LIWC-based models; Naz et al., 2025). They also offer higher generalisability and lower error rates.
- **Multimodal Methods**: The highest validity is achieved by approaches that combine text data with other sources of data, such as video or voice features. Such methods, known as multimodal, are particularly suitable for complex selection contexts such as video interviews (Hashemi-Motlagh et al., 2025).
- **Explainability and Transparency**: The traceability of AI decisions is a key success factor, especially in regulated areas such as personnel selection. Explainable models promote user trust and increase acceptance (Bhandarkar et al., 2024).

Looking at the chronological development of the individual studies analysed in the overview, a clear advancement in AI-supported methods for personality diagnostics can be observed, which is reflected in the better results of more recent work. In particular, the use of deep neural networks and multimodal architectures, for example by combining text, audio or video data, has improved the diagnostic accuracy and applicability of these methods (Naz et al., 2025; Hashemi-Motlagh et al., 2025). These methodological advances are not only reflected in the aggregated findings of meta-analyses, but are also increasingly evident in individual empirical studies investigating modern AI-based methods.

2.4 Individual Empirical Findings: Evidence for Diagnostic Added Value

While systematic reviews provide important aggregated findings on the performance of AI-supported methods, a look at individual empirical studies offers additional insights. Particularly recent work with modern language models and methodological approaches shows that AI-based analyses often enable valid, differentiated and robust assessments in real diagnostic situations. Three exemplary studies are presented below that empirically demonstrate the added value of such methods under specific conditions.

Hickman et al. (2022) show that AI-based scores from open interview responses show moderate to high agreement with classic self-ratings ($\bar{r} \approx .19$) and, in particular, with external ratings ($\bar{r} \approx .24$). The higher convergence with expert judgements suggests that machine analyses are able to deal with social desirability or self-distortion.

Sikström et al. (2025) demonstrate that text-based AI classifications of the Big Five achieve up to 10% higher accuracy than established questionnaire methods for short text inputs. At the same time, they report improved internal consistency of AI-based scales, especially for traits such as openness or neuroticism.

Another study by Tu et al. (2024) uses narcissism as an example to show that GPT-based language analyses in open-ended responses correlate significantly more strongly with expert assessments than traditional self-reports. The AI-based assessments capture relevant personality aspects more consistently and sensitively, a finding that is particularly significant for hidden or socially undesirable traits.

Overall, it is clear that AI-based diagnostic methods are not only capable of meeting classic psychometric requirements such as objectivity, reliability and validity, but can even be superior in some areas. Particularly noteworthy is the high standardisation and evaluation objectivity of algorithmic assessment, which is significantly more robust against bias than human judgements. AI-based systems are also delivering increasingly differentiated and valid results in terms of predicitive validity and behaviour assessment, especially for hard-to-access constructs such as socially desirable behaviour, subclinical personality tendencies or implicit motives. This means that they can not only complement traditional methods in a meaningful way, but also potentially surpass them in key aspects. This suggests that AI-supported tools in modern personnel selection should not be seen as a technological gimmick, but as a substantial methodological advancement.

Against this backdrop, it is worth taking a closer look at a specific practical application example: the AI-based diagnostic tool *Zortify*. The following chapter examines the extent to which this method meets scientific standards and its potential contribution to the professionalisation of modern personnel selection.

3. Zortify: Psychometric Quality of an AI-based Diagnostic Tool

The use of artificial intelligence in personality diagnostics raises legitimate questions about its scientific basis (van Iddekinge et al., 2023). The company Zortify has therefore developed a multi-stage process that combines traditional self-reports with NLP-based text analysis. The aim is to generate valid, objective and reliable personality profiles that are both theoretically and empirically sound. The following section shows

how Zortify meets the key psychometric quality criteria and what empirical evidence supports the quality of the procedure. Table 2 summarises the results.

Table 2
Evidence of the psychometric quality of Zortify

Quality criterion	Results
Objectivity	High objectivity in implementation, evaluation and interpretation thanks to fully automated processes
Internal consistency	Cronbach's α = .59–.91
Test-Retest-Reliability	r = .70 - .80 (up to 410 days)
Construct validity	Factor analysis: CFI = .98, RMSEA = .06
	Convergent validity based on HEXACO (r = .44 - .80)
Criterion validity	Individuals who received funding show significantly higher values for optimism and entrepreneurial capital, and lower values for psychopathy.
	Prediction of linguistic patterns according to scientific literature (e.g., "we" for extraversion, "I" for neuroticism)
Standardisation	Norm sample with >16,000 people from Europe

3.1 Objectivity

The objectivity of a diagnostic procedure describes the extent to which the results are independent of the person conducting the test. According to the test manual, the Zortify procedure meets this requirement to a high degree. The procedure is fully digitalised, the instructions are standardised, and the evaluation is automated using algorithm-controlled, AI-based processes. This ensures that no manual or subjective influence on the test results is possible. In particular, it is emphasised that Zortify offers objectivity in terms of implementation, evaluation and interpretation; this is a significant advantage over traditional methods that are often influenced by human judgement, such as unstructured interviews or assessment centres. Assessment centres frequently struggle with observation errors and design sensitivity, even with substantial resource investment. In contrast, AI-based methods such as Zortify demonstrate that valid diagnostics can be achieved without expensive face-to-face formats.

3.2 Reliability

High reliability is essential for personality diagnostics, as stable characteristics can only be meaningfully recorded if the assessment itself operates in a stable manner.

The empirical findings on the reliability of the Zortify method paint a convincing picture. The internal consistency of the scales, measured using Cronbach's alpha, ranges between α = .59 and α = .91, depending on the dimension. The method thus achieves values that are comparable to established questionnaire methods and meets the usual requirements for the consistency of diagnostic scales, taking into account the length of the assessment.

The temporal stability of the results was also comprehensively investigated. The correlations over periods of up to 410 days range between r = .70 and r = .80 depending on the dimension. This finding proves the long-term measurement accuracy of the method. Compared to other AI-based methods, Zortify is superior in terms of reliability. While earlier studies on text-based personality analyses by Park et al. (2015), Fan et al. (2023) or Hickman et al. (2022) report retest coefficients in the range of r = .48 to r = .70, Zortify's test-retest reliability shows higher values. This can probably be attributed to the hybrid methodology, which combines classic self-reporting with AI-supported language analysis.

Overall, the method provides a robust and temporally stable assessment of personality traits and thus meets the central requirements of a psychometrically sound instrument.

3.3 Validity

Validity is of central importance for a personality assessment tool, as it is crucial to ensure that the characteristics recorded are accurate in terms of content and interpretable in terms of meaning. The Zortify method has been systematically tested for construct and criterion validity in several empirical studies.

3.4 Construct validity

The construct validity of Zortify is supported by a series of psychometric analyses. Factor analyses confirmed the underlying structure of the method. The model quality indicators of the confirmatory factor analysis (CFI = .92; RMSEA = .06) indicate a solid model fit and support the assumption that the method represents theoretically sound personality dimensions in a representative manner.

To check the convergent validity, the Zortify method was compared with the established HEXACO inventory (N = 169). The correlations obtained underscore the substantive agreement between the two methods: for extraversion, the agreement was r = .80, and for openness, it was r = .73. Substantial correlations were also found for conscientiousness (r = .57), emotional stability (r = .43) and agreeableness (r = .44). At the same time, discriminant validity was demonstrated: the intercorrelations between scales designed to measure different constructs were low, which supports the ability of the scales to distinguish between the characteristics.

In addition, measurement equivalence across cultural contexts was demonstrated: The results show that the Zortify procedure also delivers comparable results across different European countries. This speaks for its intercultural applicability and supports the generalisability of the results in an international HR context.

3.5 Criterion validity

Criterion validity was demonstrated by several studies that examined correlations between the personality traits assessed by Zortify and external criteria such as professional performance and actual selection decisions.

A particularly practical test of predictive validity was provided by the study by Kiesow-Berger et al. (2023). In this field experiment, 812 founders applied for real funding from an external investor. All participants completed the Zortify assessment beforehand. The personality profiles of those individuals who actually received funding (n = 19) were then analysed. The results show significant differences: funded individuals scored significantly higher on optimism and entrepreneurial capital overall, and lower on subclinical psychopathy (Kiesow-Berger et al., 2023). Accordingly, these results suggest that the personality traits measured with Zortify are relevant predictors of real-world entrepreneurial funding decisions.

Further evidence of criterion validity comes from the analysis of language behaviour. In line with the scientific literature, characteristic linguistic patterns can be assigned to specific personality traits – an approach that particularly supports the validity of the NLP-based components of the method. For example, extraverted individuals have been shown to use collective pronouns such as "we" more frequently (Chen et al., 2020) This is a pattern that is also evident among Zortify users with high extraversion scores. Similarly, the expected correlation between high emotional stability and lower use of the pronoun "I" (Edwards & Holtzman, 2017) was confirmed. Openness to experience correlated as expected with the use of longer words. Machiavellian

tendencies also showed consistent language patterns in relation to pronoun use ("we"), further underlining the diagnostic validity of the semantic analysis components.

Overall, a consistent picture emerges: the Zortify method achieves a high degree of validity at both the conceptual and empirical level. It measures established personality dimensions in accordance with theory, reliably differentiates them from other constructs, shows significant correlations with external criteria, and validly maps linguistic behaviour patterns.

3.6 Standardisation

Appropriate standardisation is essential for the individual interpretation of diagnostic results. Zortify is based on an extensive European standard sample of over 16,000 participants from different countries and socio-economic contexts. This broad normative database ensures differentiated comparability of results within the European Economic Area, which is a key criterion for practical use in internationally active organisations.

3.7 Limitations and areas for improvement

Despite the convincing findings on the psychometric quality of Zortify, it is important to reflect on the methodological foundations and limitations in order to fully exploit the potential of the method and drive future developments in a targeted manner.

A comparison with established guidelines such as DIN 33430 (German Institute for Standardisation, 2021) shows that Zortify provides a solid empirical basis. For example, data on internal consistency and temporal stability over short and long intervals are available for reliability. Supplementary analyses, for example on split-half reliability or differentiation according to subgroups, can further strengthen the evidence base. With regard to validity, key aspects such as construct, discriminant and predictive validity as well as cross-cultural measurement equivalence have already been convincingly demonstrated. Studies on incremental validity and the prediction of objective performance indicators could provide further evidence.

Of particular relevance is the fundamental methodological question of how AI-supported diagnostic procedures can be adequately evaluated. In many studies, new methods are primarily validated based on their correlations with established self-report instruments. However, this approach can be problematic: if traditional methods themselves have methodological weaknesses, such as social desirability or limited

objectivity, there is a risk that innovative methods such as Zortify will be systematically underestimated. This problem is also highlighted in recent meta-analyses (e.g., Naz et al., 2025; Bhandarkar et al., 2024), which argue for the development of new validation standards for AI-based diagnostics. Possible options include more criteria- and behaviour-oriented validations or the development of convergent evidence across different data sources (e.g. language, behaviour). At the regulatory level, for example in the context of the EU AI Act, it could be specified in greater detail in future how psychological AI systems should be tested in order to meet both scientific and ethical requirements.

Overall, Zortify represents an innovative, empirically sound and standardised approach to personality diagnostics. The validations carried out to date demonstrate high psychometric quality. At the same time, the technological and methodological landscape is evolving strongly, making ongoing research, openness to further developments and discussion of appropriate evaluation criteria essential in order to fully realizing and responsibly utilising the potential of AI-based diagnostics in the long term.

4. Summary and Outlook: Why Personnel Selection Needs Evidence-Based AI Now

Despite the long-recognised limitations of traditional selection selection methods and the well-documented fallibility of human judgement, personnel selection in practice remains largely resistant to evidence-based methods and innovations. Methods with insufficient psychometric properties, such as unstructured interviews or the free interpretation of CVs, continue to dominate the selection process. This status quo is rarely due to ignorance, but more often to institutional inertia, pragmatic considerations and a persistent underestimation of the role of systematic biases in human judgement.

At the same time, a growing number of empirical studies show that modern, AI-based methods, especially those based on natural language processing (NLP), have the potential to substantially improve the quality of selection decisions. They can provide more differentiated, valid, and less biased methods for assessing personality than traditional questionnaires or interviews. This makes them particularly valuable in contexts with a high risk of bias or when certain characteristics are difficult to measure. The psychometric evaluation of the Zortify method provides a practical example of this: reliability and validity reach a level that corresponds to conventional

questionnaire instruments based on classical test theory and also supplements an indirect behavioural component that cannot be specifically manipulated. Zortify shows better psychometric quality than all AI instruments published to date. In addition, evidence-based AI methods like Zortify enable fully automated, location-independent implementation with minimal resource requirements. They offer a sustainable alternative to traditional, resource-intensive selection procedures, particularly with respect to scalability and long-term cost reduction.

Nevertheless, the use of AI in personnel selection is not an end in itself. The principle of of context-dependence remains central: not every AI application is useful, nor is every application valid. Research shows that methods that are psychologically sound, empirically tested and transparent in their functioning are particularly successful. Ethical standards, data protection guidelines and regulatory requirements must also be consistently considered and implemented.

A sustainable approach to AI in personnel selection therefore requires two things: First, a willingness to take evidence-based procedures seriously and turn your back on traditional routines. Assessment centres are a prime example of the gap between traditional practice and empirical evidence: they are costly, difficult to standardise and, despite their high cost, often associated with questionable diagnostic quality. The fact that they are nevertheless widespread demonstrates the inertia of diagnostic routines. Secondly, it requires a responsible approach to technological innovation: AI procedures must not be adopted uncritically, but must be scientifically tested, used in a context-sensitive manner and designed in a legally responsible way.

AI-based diagnostics can make personnel selection fairer, more informed and more effective. Given the susceptibility of traditional methods to bias and the numerous advantages of AI-based diagnostics, it is not only wise but essential to use modern, evidence-based AI methods today.

Heidbrink, M., & Feltes, F. (2025). *From Intuition to Algorithm: Why AI is Becoming Indispensable in Personnel Selection* [White paper]. Zortify. https://zortify.com

References

Armoneit, C., Schuler, H., & Hell, B. (2020). Nutzung, Validität, Praktikabilität und Akzeptanz psychologischer Personalauswahlverfahren in Deutschland 1985, 1993, 2007, 2020: Fortführung einer Trendstudie. *Zeitschrift für Arbeits- und Organisationspsychologie*, *64*(2), 67–82. https://doi.org/10.1026/0932-4089/a000311

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*(4), 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Bhandarkar, A., Wilson, R., Swarup, A., Webster, G. D., & Woodard., D. (2024). Bridging minds and machines: Unmasking the limits in text-based automatic personality recognition for enhanced psychology-AI synergy. *British Journal of Psychology, 00, 1-23.* https://doi.org/10.1111/bjop.12755

Chen, J., Qiu, L., & Ho, M.-H. R. (2020). A Meta-analysis of linguistic markers of extraversion: positive emotion and social process words. *Journal of Research in Personality, 89*, 104035. https://doi.org/10.1016/j.jrp.2020.104035

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*(3), 493-511. https://doi.org/10.1037/0021-9010.72.3.493

Hashemi-Motlagh, S. M., Rezvani, M. H., Khounsiavash, M. (2025). AI methods for personality traits recognition: a systematic review. *Neurocomputing, 640,* 130301. https://doi.org/10.1016/j.neucom.2025.130301

Deutsches Institut für Normung. (2016). *DIN 33430: Eignungsdiagnostik.* Beuth Verlag.
Edwards, T., & Holtzmann, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality, 68*, 63-68. http://dx.doi.org/10.1016/j.jrp.2017.02.005

Fan, Y., Zhao, Y., Le, T. Q., Srikant, S. A., Fu, J., & Chen, Y. (2023). How well can an AI chatbot infer personality? A psychometric evaluation of language-based personality estimation. *Proceedings of the National Academy of Sciences, 120*(30), e2302337120. https://doi.org/10.1037/apl0001082

Hickman, L., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology, 107*(2), 244–262. https://dx.doi.org/10.1037/apl0000695

Kiesow-Berger, H., Fell, M., Philippy, F., & Steiner, E. (2023). KI-Unterstützung bei der digitalen Personalauswahl – Eine Fallstudie im Unternehmer-Kontext. In P. Stulle & R. T. Justenhoven (Hrsg.) *Personalauswahl 4.0* (pp. 163-177). Springer Gabler.

Koutsoumpis, G., Leicht-Deobald, U., & Li, C. (2024). Beyond traditional interviews: Psychometric analysis of asynchronous video interviews (AVIs). *Journal of Applied Psychology*. Advance online publication. https://doi.org/10.1016/j.chb.2023.108128

Kowalski, C. M., Rogoza, R., Vernon, P. A., & Schermer, J. A. (2018). The Dark Triad and the self-presentation variables of socially desirable responding and self-monitoring. *Personality and Individual Differences, 120*, 234–237. https://doi.org/10.1016/j.paid.2017.09.007

Moreno, J. D., Martínez-Huertas, J. Á., Olmos, R., & García-Sancho, E. (2021). A meta-analysis of the validity of text-based personality detection using natural language processing. *Journal of Research in Personality, 93*, 104130. https://doi.org/10.1016/j.paid.2021.110818

Naz, A., Khan, H. U., Bukhari, A., Alshemaimri, B., Daud, A., & Ramzan, M. (2025). Machine and deep learning for personality traits detection: a comprehensive survey and open research challenges. *Artificial Intelligence Review, 58*, 239. https://doi.org/10.1007/s10462-025-11245-3

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology, 108*(6), 934–952. https://doi.org/10.1037/pspp0000020

Rahe, C., & Rahe, T. M. (2017). *Nutzung von personaldiagnostischen Instrumenten bei der Auswahl von Führungskräften in deutschen Unternehmen*. Rahe Management Consultants.

Sackett, N., Zhang, P. R. Berry, C. M., & Lievens, F. (2022). Revisisting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology, 107*(11), 2040-2068. https://doi.org/10.1037/apl0000994

Sikström, S., Valavičiūtė, I., & Kajonius, P. J. (2025). Personality in just a few words: Assessment using natural language processing. *Personality and Individual Differences, 238*, 113078. https://doi.org/10.1016/j.paid.2025.113078

Tu, Z., Zhang, Z., Zhang, W., Luo, F., & Bian, R. (2024). Using large language models to identify narcissism based on texts. *PsyArXiv*. https://doi.org/10.2139/ssrn.4965442

Van Iddekinge, C., Lievens, F., & Sackett, P. R. (2023). Personnel Selection: A review of ways to maximize validity, diversity, and the applicant experience. *Personnel Psychology 76*(2), 651-686. https://doi.org/10.1111/peps.12578

Yarkoni, T. (2010). Personalit in 100,000 words: a large scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 44*, 363-373. https://doi.org/10.1016/j.jrp.2010.04.001

Zell, E., & Lesick, T. L. (2022). Big Five personality traits and performance: A quantitative synthesis of 50+ meta-analyses. *Journal of Personality, 90*(4), 559–573. https://doi.org/10.1111/jopy.12683

Zhang, C., Lu, Z., & Wang, M. (2024). Can Large Language Models assess personality from asynchronous video interviews? A zero-shot evaluation of GPT-3.5 and GPT-4. *PsyArXiv Preprint*. https://doi.org/10.1109/TAFFC.2024.3374875