



ARTICLE



<https://doi.org/10.1057/s41599-026-09131-6>

OPEN

Designing fairness: best practices for gender-sensitive development of cognitive ability tests in recruitment

Jannick Schneider^{1,2}, Clemens Striebing¹, Melanie Elizabeth Jacobsen³ & Katja Pässler⁴

Psychological tests play a pivotal role in high-stakes decisions such as recruitment, yet traditional development guidelines concentrate fairness work downstream—at test administration and post hoc statistical bias correction—while offering little concrete guidance for the design stage, where constructs are defined and operationalized. Drawing on constructivist and organizational justice theories, we argue that a “bias by design” arises when unexamined default assumptions narrow construct representation, systematically disadvantaging groups before any item is administered. We develop this for cognitive ability tests, which exhibit the largest subgroup differences among common selection instruments and routinely rely on figural matrices as proxies for general ability, thereby disadvantaging women. We contrast procedural justice (standardized administration) with distributive justice (equitable score distributions), clarifying that the latter applies to a principled class of constructs—latent and content-general, such as fluid intelligence—for which content-linked group differences signal construct-irrelevant variance rather than true differences. To address this gap, we adapt Stanford’s Gendered Innovations framework and demonstrate its application through the Modularer Kurzintelligenztest (M-KIT; Dantlgraber et al., 2015). Fairness can be addressed at three hierarchically ordered levels, where higher levels constrain lower ones: the theoretical model (broadening construct representation), the task format (removing construct-irrelevant demands), and the item (differential item functioning as final refinement, not primary remedy). We distill practical recommendations for embedding distributive justice early, including questioning default assumptions and documenting fairness deliberations. Our approach reframes fairness not as a trade-off but as integral to validity, offering a roadmap for more inclusive assessments.

¹Fraunhofer Institute for Industrial Engineering, Stuttgart, Germany. ²Ruhr University Bochum, Bochum, Germany. ³Weizenbaum Institute for the Networked Society, Berlin, Germany. ⁴FHNW University of Applied Sciences and Arts Northwestern Switzerland School of Applied Psychology, Olten, Switzerland.
✉email: jannick.schneider@iao.fraunhofer.de; melanie.jacobsen@weizenbaum-institut.de

Introduction

“The fact is that gender ideology does not operate as an explicit force in the construction of scientific theories. Its impact (like that of ideology in general) is always indirect: in the formation and selection of preferred goals, values, methodologies, and explanations” (Keller, 1995, p. 137).

Psychological tests are widely used to inform recruitment, educational, and other aptitude-related decisions. For example, tests of cognitive ability (Arnoneit et al., 2020; Kulikowski et al., 2025), situational judgement tests (Cabrera and Nguyen, 2001), or achievement motivation tests (Schmidt-Atzert, 2004) could indicate whether a person fits into a certain degree program or job position. Indeed, test scores in the past have been documented to play a crucial role in predicting important life outcomes such as academic (Rohde and Thompson, 2007) and job performance (Schmidt and Hunter, 1998; Steel and Faroborzi, 2024). Thus, because of the wide-reaching consequences a test-based decision could have for an applicant's or student's future, it is important to ensure that no group of individuals systematically scores higher or lower than other groups for reasons that cannot be reasonably justified – that is, to ensure the criterion of test fairness is fulfilled. Not surprisingly, there are lively discussions about fairness regarding an applicant's cultural background (Reynolds and Suzuki, 2012; Stumpf et al., 2017; Wilson et al., 2023), ethnicity (Chung-Yan and Cronshaw, 2002; Reynolds et al., 2021; Schmidt and Hunter, 1974), and gender (Herzberg, 2024; Sparfeldt and Schult, 2024; Striebing et al., 2024).

The guidelines typically used as a basis for developing psychological tests – and, still more so, the degree to which developers adhere to them – concentrate fairness work at the stages of test administration and post-hoc statistical evaluation, while offering comparatively little concrete guidance for the design stage; even those few recommendations that do apply to design are followed inconsistently in practice (e.g., Jonson et al., 2019). In this paper, we argue that different conceptions of fairness play a crucial role in producing this asymmetry. Where emphasis is placed on the test criterion of objectivity, attention gravitates toward procedural justice, that is, toward process-related aspects of testing. We propose that greater attention should be given to distributive justice—how test results are allocated across subgroups—and, crucially, that this attention must be paid at the point where constructs are defined and operationalized in recruitment settings. The structural problem with unfairness in aptitude tests therefore does not lie solely in the administration of psychological tests (e.g., standardization) or in statistical considerations after testing has concluded (e.g., differential item functioning; DIF). These steps are necessary, but by the time they are applied, the decisive choices—which facets of a construct are represented, and through which stimulus formats—have already been fixed. Unfairness, in this sense, is not primarily a failure of correction but a consequence of the basic assumptions underlying test development, and hence, a bias by design. Because cognitive ability tests are among the most widely used instruments for aptitude-related decisions (Arnoneit et al., 2020; Pollard et al., 2015), they are the exclusive focus of this paper. The restriction is deliberate: of the instruments commonly used in selection, cognitive ability tests exhibit the largest and most consequential subgroup differences (Woods and Patterson, 2024), and within them figural matrices are routinely employed as efficient proxies for general cognitive ability—which makes the consequences of a narrow operationalization particularly visible. This concern is not confined to the research literature. Short cognitive ability tests are increasingly administered at scale by commercial test providers and consulting firms, typically as online screening instruments applied early in the selection funnel and before any human

review, yet many such instruments are of uncertain psychometric quality and are accompanied by little or no published validity evidence (Hossiep et al., 2015; Prochazka et al., 2023). A narrow operationalization therefore does not merely distort a single score; it is replicated across large applicant pools at the point where the majority of candidates are screened out.

Using gender as a case in point, we demonstrate in this article that underlying assumptions in test development reflect different notions of fairness, resulting in diverging priorities and levels of sensitivity. We then present a guide to gender-sensitive test development based on the case study of a cognitive ability test that actively considered gender fairness in its theory, development, and design (Dantlgraber et al., 2015). Thus, this test constitutes one of the first test manuals addressing gender fairness as a design paradigm. We base our recommendations for developing gender fair tests on the Gendered Innovations approach by Stanford University (Schiebinger et al., 2011–2020). It provides impulses for considering gender in every step of a scientific process and, thus, enhances previous recommendations for fairness in test development (Jonson et al., 2019) through its constructivist¹ perspective on the basic assumptions of research and development (R&D) processes.

To illustrate the need for considering gender fairness in the test construction phase, we start by providing an example of how a lack of theoretical soundness in test design might result in questionable test scores and therefore potential bias in decision-making.

Bias by design in aptitude tests

Psychological tests are defined by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education in the following way: „A test is a device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process” (AERA et al., 2014, p. 2). Utilizing such tests in recruiting processes, recruiters aim at identifying skills, knowledge, abilities, or other more generic characteristics of applicants (e.g., cognitive abilities, personality characteristics or motives and interests; Farr and Tippins, 2017; Höft et al., 2017). These should in some way relate to the relevant hiring criterion (often job performance) to help recruiters selecting those applicants who are most suitable for the specific job (Arnoneit et al., 2020; Murphy, 2017).

Striebing and colleagues (2024) argue that women are systematically disadvantaged in cognitive ability tests because these tests regularly prioritize types of cognitive abilities in which men tend to score higher, namely mathematical and non-verbal reasoning abilities (Waschl and Burns, 2020). Thus, the lack of a comprehensive construct representation—such as assessing inductive reasoning based on numerical, figural, and alphabetical stimuli (Jäger et al., 1997)—can lead to a gender biased conceptualization and assessment in application settings (Messick, 1989, 1995). This is particularly problematic given that general intelligence is roughly equally distributed between men and women (Flynn and Rossi-Casé, 2011).

This form of gender-specific test design may be justified if the relevant skills, which are distributed unevenly between the genders, are clearly and comprehensibly important specific professional skills of the position in question. However, an optimum match between the skills required by the job and the aptitude tests used is not practiced regularly. Time and cost constraints in job assessment settings require efficient, compact, and easy-to-use measures of applicants' knowledge, skills, and abilities. Because of

the required efficiency in personnel selection, the range of cognitive abilities assessed is often only focused on a single dimension (e.g., figural inductive reasoning; Formann and Pischwanger, 1979; Hossiep and Hasella, 1999; Raven, 1939; Weiß, 2019). Intelligence is multidimensional, as are its subcomponents such as inductive reasoning, working memory capacity, and processing speed (Dantlgraber, 2015; Jäger et al., 1997). Tests that omit some of these facets sacrifice theoretical and stimulus comprehensiveness, which can bias the prediction of criteria such as job performance across groups (Waschl and Burns, 2020).

The described phenomenon is, in psychometric terms, an instance of construct underrepresentation—one of the two principal threats to validity in Messick's (1989, 1995) unified framework alongside construct-irrelevant variance. When a test intended to measure inductive reasoning samples only figural content, it undersamples the construct domain, and the resulting score reflects the construct less faithfully than its label implies. This threat is well established. Hence, our argument is narrower and distinct in three respects. First, the validity literature treats underrepresentation as a property to be detected in a finished instrument but offers no account of why specific underrepresentations recur and why they fall systematically along gender lines. Second, underrepresentation is a validity problem irrespective of group impact; we are concerned specifically with the subset that produces systematic group disadvantage—what we term bias by design. Third, although design-stage validity frameworks exist (e.g., Evidence-Centered Design; Mislevy and Haertel, 2006), they operationalize coverage of the construct as defined rather than interrogating whether the construct definition itself encodes an unexamined default. It is this last step that the Gendered Innovations approach (Schiebinger et al., 2011–2020) is designed to address.

A typical example of a gender-unequal cognitive ability assessment that might lead to gender imbalanced results is inductive reasoning. Inductive reasoning represents the ability to discover underlying rules in relationships or information and to apply these rules to obtain a solution (Carroll, 1993). It is disproportionately assessed with figural tasks, thus assessing figural inductive reasoning, neglecting other types of stimuli (e.g., verbal/alphabetical or numerical; Wasch and Burns, 2020). This is illustrated in tasks like figural matrices. These tasks load highly on fluid intelligence (Gignac, 2015) and are therefore applied frequently as an efficient measure of cognitive ability in occupational settings. Test-takers are required to identify underlying rules within a visual pattern (the matrix) and select the correct solution from a set of response options (distractor-based format; Koch et al., 2022). A meta-analysis by Lynn and Irwing (2004) found a male advantage of five IQ points in figural progressive matrices tasks as developed by Raven (1939). This variance disappeared when evaluating a general factor of fluid intelligence (Giofrè et al., 2022).

One reason for explaining these differences in test performance is a contaminated assessment of inductive reasoning when measured through figural tasks, as the figural nature of these items requires individuals to additionally rely on visuo-spatial abilities (Waschl and Burns, 2020), especially with increased complexity of the displayed figures (Arendasy and Sommer, 2012). For this reason, we argue that there is potential bias by design in the construction of efficient cognitive ability tests for application settings that may lead to unfair group differences. Administering such a seemingly standardized test to a group consisting of both men and women could therefore lead to women being unjustifiably singled out because construct-unrelated factors influence test performance, even though they are actually well-suited candidates (Azen et al., 2002; Cole and Zieky, 2001; Woods and Patterson, 2024).

But why are such biased tests widely practiced, even though, on closer inspection, they do not meet the quality standards for psychological testing, besides the above-presented resource efficacy in applied settings? In the following, we argue that their continued use is due to the way test quality standards and the concept of fairness are interpreted and prioritized by test developers and test service providers. In this article, we do not intend to discuss the best way to determine and eliminate test bias after the testing procedure (e.g., Bazian et al., 2025) using statistical methods (for an overview, see Aguinis and Culpepper, 2024). We want to discuss how fairness can be established from the much earlier stage of test development, thus tackling bias by design.

Distributive and procedural fairness. The concept of fairness can be related to perceptions of so-called procedural and distributive justice (Gilliland, 1994).² In the context of organizational justice literature, procedural justice denotes the perceived fairness of formal procedures and mechanisms through which decisions are reached (Greenberg and Tyler, 1987). Aspects related to formal characteristics of the process, adequate explanation of the procedure, and interpersonal treatment contribute to an individual's perception of procedural justice (Greenberg, 1990).

Beyond the common emphasis on subjective perceptions of fairness, definitions of procedural justice can also include a normative perspective, which was introduced for psychological testing. In this sense, Kane (2010) stated that procedural fairness in psychological testing requires that “all test takers be treated in essentially the same way, that they take the same test or equivalent tests, under the same conditions or equivalent conditions, and that their performances be evaluated using the same (or essentially the same) rules and procedures” (Kane, 2010, p. 178). While procedural fairness is considered an important aspect of psychological testing, there are clear parallels to the testing criterion of objectivity—meaning that the testing process is reproducible independently of the place, time, test administrator, and test evaluator (AERA et al., 2014; Bühner et al., 2017; Moosbrugger and Kelava, 2020)—which raises the question of whether the scope of this conceptualization of fairness may be too narrow, given that reproducibility and standardization alone do not ensure fairness.

Distributive justice, by comparison, relates to an individual's perception that the effort they put into an organizational process is proportionally reflected in their own organizational outcome, especially compared to the input-output ratio of another individual of the same organization (Greenberg, 1990). Within cognitive ability testing, the outcome corresponds to the test result, which can be expected to be equal across genders if the testing situation is equal for both groups and the same effort and abilities are put into the test. Referring to our previous example on cognitive ability and inductive reasoning tests: If two applicants, categorized within a dichotomous gender framework (man and woman), possess the same level of inductive reasoning but the test assesses it solely through figural matrices, the design may be biased (for example, as visuo-spatial abilities confound the ability assessment and construct accessibility). Relying exclusively on figural inductive reasoning to evaluate overall inductive ability and to predict job performance (so-called construct underrepresentation) can disadvantage women, reducing distributive justice, as suggested by empirical findings above.

Considering that such starkly simplified conceptions of cognitive abilities are the case, it is surprising that no definition of a concept such as distributive fairness in psychological testing has been included in any official guidelines issued by leading professional associations or psychometric standardization bodies

(e.g., AERA, APA, Deutsches Institut für Normung [DIN]). Indeed, we argue that distributive fairness has received insufficient attention in psychological testing.

Distributive fairness in the sense we intend applies to a specific and principled class: constructs that are latent and content-general—such that content-specific variance is, by definition, not part of the construct—and for which there is independent theoretical and empirical warrant to expect no true group difference at the level of the core construct. Fluid intelligence meets both conditions: it is defined as a content-general information-processing capacity (Jensen, 1998; McGrew, 2009), and general cognitive ability shows negligible sex differences (Colom et al., 2000; Giofrè et al., 2022). So, when verbal, numerical, and figural tasks all load on the same latent factor but only the figural format produces a sizable group gap, content-linked variance is construct-irrelevant by definition (Campbell and Fiske, 1959). Constructs whose group differences are constitutive rather than artifactual fall outside this scope. Reaction time is instructive here: it differs reliably by age, that difference is part of what the construct measures, and it cannot be made to appear or disappear by re-selecting stimulus content (Der and Deary, 2009). For such constructs, equalizing outcomes would compensate for genuine differences, which is neither our claim nor a defensible one. The distinction is therefore between differences that are constitutive of the construct and differences introduced by a discretionary feature of its operationalization (see below how the M-KIT approached this issue).

In the context of psychological testing, test fairness has traditionally been placed apart from the classical quality criteria of objectivity, reliability, and validity, listed as a further criterion rather than among the core. This positioning shifted in 2014, when the *Standards for Educational and Psychological Testing* (AERA et al., 2014) treated fairness as foundational to test validity in the US context, a development not yet mirrored in, for example, the German-speaking assessment tradition. The AERA defines fairness based on procedural aspects (“fair and equitable treatment of all test takers”) and fairness as the absence of measurement bias, construct accessibility, and valid score interpretation (AERA et al., 2014, p. 51).

We do not claim that fairness is neglected in contemporary psychometrics. On the contrary, the current *Standards* treat it as foundational, defining bias in terms of construct-irrelevant variance (AERA et al., 2014). Our claim concerns one central asymmetry, whose reach is larger than its usual framing suggests: the operationalization of fairness remains concentrated downstream. Fairness evidence is elaborated for measurement bias, response processes, and consequences of use—all presupposing a fixed construct definition—whereas procedures for interrogating that definition are underdeveloped (Jonson et al., 2019). The same concentration marks the standard most specific to our context, the *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2018), which treat fairness chiefly as equitable treatment, comparable access, and the absence of predictive and measurement bias, while interrogation of the construct definition falls under validity rather than being a fairness decision point.

The *Principles*, following the *Standards*, reject equality of group outcomes as a definition of fairness and hold that group differences do not in themselves indicate bias. We do not dispute this; our claim is not that outcome differences are bias, but that they warrant scrutiny of the operationalization—which follows the *Principles*’ own logic that group differences should trigger heightened scrutiny for construct underrepresentation and construct-irrelevant components (SIOP, 2018, p. 22). On a latent, content-general construct, a difference confined to a single stimulus format is precisely such a flag (see Recommendation 8). Our proposal thus extends, rather than contradicts, the scrutiny

these standards prescribe, relocating it to the design stage, where the operationalization is still open to revision. The asymmetry thereby proves decisive: the stage at which construct-irrelevant variance can still be prevented is the one least served by existing fairness procedures—and it is most pronounced in the German-speaking standards (DIN, 2016; Bühner et al., 2017; Moosbrugger and Kelava, 2020), where fairness remains subordinated to objectivity. Our contribution addresses this asymmetry within the context most affected by it.

Given our above example on inductive reasoning as a go-to measure for cognitive abilities in personnel assessment, valid score interpretation is difficult when the measurement of inductive reasoning is confounded through construct-unrelated influences. As Kane (2010) stated: “An assessment that is unfair, in the sense that it systematically misrepresents the standing of some individuals or some groups of individuals on the construct being measured or that tends to make inappropriate decisions for individuals or groups is, to that extent, not valid for that interpretation or use. Similarly, an assessment that is not valid in the sense that it tends to generate misleading conclusions or inappropriate decisions for some individuals or groups will also be unfair.” (Kane, 2010, p. 181). This leads to our next question: How are fairness and validity of psychological tests related?

Trading fairness for validity? In psychological testing and personnel selection, the most heavily weighted test criterion for informing decision-making is the validity of the selection procedures—particularly their predictive, criterion-related validity. Previously, various studies discussed a trade-off between this type of validity and fairness (Speer et al., 2025). The following section provides a brief overview of recent discussions addressing this potential trade-off.

As the intended purpose of psychological tests is to inform aptitude-related decisions, ideally, the test accurately predicts the test taker’s performance in the organization they are selected for. Therefore, if the correlation between the test score and later performance is high, the test score is said to have high predictive validity, and its use is encouraged. For example, in the case of cognitive ability tests in employee recruitment, their predictive validity was said to be high (Schmidt and Hunter, 1998). Consequently, they have been popular tools for selecting employees (Armoneit et al., 2020; Pollard et al., 2015). Considering that high predictive validity often coincides with a lower pass rate of minorities and women, researchers have referred to this as the so-called diversity-validity dilemma (Ployhart and Holtz, 2008; Pyburn et al., 2008).

However, not only has the overreliance on predictive validity estimates for the application of psychological tests been criticized (e.g., Borsboom, 2006); a recent comprehensive meta-analysis by Sackett and colleagues (2024) resulted in predictive validity estimates that were considerably lower for cognitive ability tests than previously reported. It has even been proposed that cognitive ability tests could be left out of the recruiting process entirely to reduce adverse impact (Berry et al., 2024; Sackett et al., 2024). Thus, the so-called diversity-validity dilemma stops being a dilemma if validity’s weight in the equation is greatly reduced (Sackett et al., 2023), and it might make more sense to prioritize fairness when developing and applying cognitive ability tests.

Furthermore, research supports that most tests used for employee recruitment in practice are outdated, of low psychometric quality, or developed not by experts in psychometrics, but by consulting firms that provide little to no validity estimates (Hossiep et al., 2015; Prochazka et al., 2023). One can assume that these tests suffer from the same lack of comprehensive construct representation as previously described. The quality of many of the

tests commonly used in practice is, thus, hard to evaluate and necessitates a bigger focus on test fairness.

The gender bias by design described here is sometimes justified on the grounds that it is fair, in the sense that although the respective test design may tend to lead to lower chances of success for certain social groups, it would still identify the best performers overall (e.g., Herzberg, 2024; Sparfeldt and Schult, 2024). This must be countered by the fact that this performance equity paradigm of psychological tests is being viewed increasingly critically (Dantlgraber et al., 2015; Ployhart und Holtz, 2008; Woods and Patterson, 2024).

In modern psychological testing situations, it is common to test for so-called adverse impact (Hough et al., 2001) or “the effect by which the proportion of one group meeting a selection criterion is systematically different from another” (Woods and Patterson, 2024, p. 256). A recent meta-analysis revealed that the exclusion of cognitive ability tests results in only a negligible decline in predictive validity, while substantially reducing adverse impact (Berry et al., 2024). If human resource professionals still choose to administer cognitive ability tests during assessments, they can at the very least ensure that the measures used are fair. In the following, we will discuss past recommendations on how this fairness can be addressed in test development and administration.

The state of fairness guidance in current standards

Even though fairness is an essential component in the updated Standards for educational and psychological testing (see above), as “fairness is a central issue in achieving valid test results” (Plake and Wise, 2014, p. 7), prevalent measures still perform poorly under these standards and lack adherence to established recommendations (Jonson et al., 2019). Additionally, many of the currently established instruments were developed before these standards became widely known to researchers, thus making compliance difficult.

Since our discussion in this paper thus far has primarily been theoretical and conceptual, we now discuss the general lack of practical recommendations addressing fairness in test development within past guidelines. We focus on the development stage, as strategies to reduce bias after test items are administered are less effective (Bazian et al., 2025).

Lack of practical recommendations for developing fair psychological tests. Considering current conceptualizations of fairness, Xi (2010) remarked that “previous fairness frameworks (...) may not provide practical guidance on how to go about developing the relevant evidence to support fairness” (Xi, 2010, p. 148). One comprehensive study was conducted by Jonson and colleagues (2019), who collected 17 practical recommendations from established psychological testing guidelines and reviewed their consideration in established general intelligence and cognitive ability tests. Not only did most tests exhibit a lack of compliance with the recommendations, but only three out of the 17 recommendations were applicable during the test design phase (*use of universal design principles in test design and development, review of test by a panel with subgroup expertise, describing test review process and results*). The remaining criteria focused on subgroup analysis after test administration as well as documentation of the testing process, mirroring a focus on analytical and procedural aspects of testing (see procedural justice above). Thus, there is a need for more practical guidelines for the development of fair psychological tests in recruitment settings, as previous recommendations might not cover this aspect in depth based on previous elaborations for potential bias by design.

One interdisciplinary approach to enhance fairness considerations in test design can be the Gendered Innovations framework

(Schiebinger, 2011–2020), which has been applied to various scientific sub-disciplines with the goal of fostering scientific discovery and improving social equality (Tannenbaum et al., 2019). In the following, we will provide a brief description of the idea of Gendered Innovations and how these can be related to psychological testing.

The gendered innovations approach. The Gendered Innovations approach (Schiebinger et al., 2011–2020; Tannenbaum et al., 2019) aims to systematically integrate sex and gender analysis into scientific and technological research. This approach recognizes that overlooking gender differences can lead to inaccuracies, inefficiencies, and even inequities in various domains, including psychological testing. The methodology promotes interdisciplinary collaboration and the adoption of best practices to identify and mitigate gender biases at all stages of research and development. Based on a social-constructionist perspective, the Gendered Innovations approach assumes that scientists engage in R&D processes based on their own specific presumptions and perspectives. These presumptions, when not reflected and balanced, influence outcomes and potentially gender-based differences in developed services or products. Hence, we argue that psychological test construction for personnel assessment can benefit from questioning presumptions early in the process of test development by reflecting on the underlying theoretical foundation (e.g., why are figural matrices so widespread, and how might this influence fairness? We discuss this below).

Key aspects of the Gendered Innovations approach include the early incorporation of gender considerations in research designs, the disaggregation of data by sex and gender, and the development of gender-responsive methodologies (Tannenbaum et al., 2019). Rather than treating gender as an afterthought or applying corrective measures post hoc, this approach emphasizes that integrating gender-sensitive analyses from the outset leads to more robust, valid, and fair outcomes.

In the context of psychological testing for personnel selection, the Gendered Innovations approach can provide a structured way to identify and address biases embedded in test design. In line with the approach of engineering innovation processes (Schiebinger et al., 2011–2020), the development of a new cognitive ability measure can be conceptualized as the development of a new product or service to integrate in hiring processes (Farr and Tippins, 2017; Höft et al., 2017). The test developers, in their role as designers, base their approach to this process on their external knowledge (acquired in training) and internal knowledge (acquired in the specific workplace), as well as their experience of success and impact of the products and services developed.

As such, test developers can consider fairness and gender aspects in different steps of the development process. Starting with rethinking research priorities and outcomes, theories, research questions, and the inclusion of different perspectives in the process, test developers can reflect upon the knowledge and assumptions embedded in these initial steps. As these steps lay the groundwork for the entire research process, the fairness of R&D processes depends primarily on these initial considerations. In further steps of the process, researchers can analyze their data by sex, gender, and intersecting factors potentially threatening the fairness for other socio-demographic groups. Finally, the process encourages rethinking standards and reference models (see above) for future research approaches.

When comparing this R&D process to past guidelines and practical recommendations (Jonson et al., 2019), psychological test development and fairness standards are quite elaborate when it comes to the analysis and reporting of sub-group differences. Statistical tools such as measurement invariance testing (Leitgöb

et al., 2023), DIF (Bauer, 2017; Wetzel and Hell, 2013), predictive bias analysis (Aguinis and Culpepper, 2024), and adverse impact analysis for personnel selection (Woods and Patterson, 2024) help researchers identify potential threats to fairness after test administration. However, none of these procedures can flag that a construct was defined too narrowly in the first place, because each operates on an already-fixed operationalization. Remedies applied at this stage therefore address a symptom rather than its cause: post-hoc item selection cannot repair a gap that originates in the operationalization (Bazian et al., 2025), and only a minority of fairness recommendations reviewed by Jonson et al. (2019) apply at the design stage at all (see above). Group norms illustrate the same limitation and add a further caveat, as they adjust the score distribution without touching the instrument. If the group difference reflects construct-irrelevant variance, norming masks a measurement artifact rather than removing it; if instead it reflects a genuine difference, it equates scores by group membership, which in selection contexts threatens equal treatment.

The Gendered Innovations approach, on the other hand, enhances previous guidelines on test development in the first part of the process: rethinking research priorities, outcomes, concepts, and theories. Encouraging test developers to reflect on their underlying assumptions promotes a more comprehensive approach to fairness. This includes examining whether test formats, such as those used in cognitive ability tests (e.g., figural matrices), unintentionally favor one gender and thus lead to biased outcomes (Waschl and Burns, 2020). The following section translates key principles of the Gendered Innovations approach to psychological testing.

Rethinking concepts and theories: gendered innovations in psychological test development. Assuming concepts by default can lead to further pigeonholing research: “Default means failure to act; neglect or a preselected option adopted when no alternative is specified” (Schiebinger, 2014, p. 5). As mentioned above, quality criteria of psychological test development refer to procedural conceptualizations of justice, lacking a more elaborated perspective on fairness issues in modern testing environments (e.g., personnel assessment and the representation of different genders or minorities). The AERA guidelines replace objectivity with fairness, thus assigning higher priority to various issues of fairness in test development and further elaborate on aspects of distributive justice (e.g., fairness as the absence of measurement bias, construct accessibility, and valid score interpretation).

Using the example of inductive reasoning tests provided previously, the use of mostly figural tasks to assess inductive reasoning is rooted in the assumption that these tasks are equally known or unknown to all test takers (Waschl and Burns, 2020). Questioning this assumption paves the way for developing alternative tasks that are not solely based on figural skills and consequently represent the construct of inductive reasoning more comprehensively (Dantlgraber, 2015; Waschl and Burns, 2020). As a result, a more valid score interpretation of applicants’ inductive reasoning might be achieved, as multiple perspectives are considered. Hence, new formats of items to assess the underlying psychological construct might be required to enhance fairness in testing inductive reasoning.

The Gendered Innovations approach integrates (gender) fairness considerably earlier in the test development process than is usually recommended, for instance by the *Standards for Educational and Psychological Testing* or the guidelines compiled by Jonson and colleagues (2019). By rethinking concepts and theories, test developers are encouraged to reflect on their basic assumptions and the underlying theory (e.g., theories of general intelligence, or of its development in childhood and adolescence),

and on how much emphasis they place on procedural as opposed to distributive aspects of justice. This reflection extends to their core concepts (e.g., fluid intelligence and inductive reasoning), which in turn shape the hypotheses developers hold about what their test material should yield in terms of distributive fairness. Only on this basis are the methodology and test material developed and the resulting data analyzed. The premise that the initial steps of the research cycle deserve this scrutiny is not unique to Gendered Innovations: it converges with recent proposals for participatory and community-engaged psychological research, which likewise locate the decisive choices at the stage of conceptualization and design and ask researchers to examine the assumptions and positionality they bring to it, rather than addressing representational concerns only at the stage of analysis or reporting (Mohamed et al., 2026).

In the second part of this article, we integrate this call for action in the field of fair psychological testing and develop best practices (Table 1) for gender-sensitive test development based on the reviewed literature and a measure of cognitive abilities developed with gender fairness and resource constraints of application settings in mind (Dantlgraber et al., 2015). This case study can be considered as a contextualization of the Gendered Innovations approach to fairness in psychological testing, thus enriching the standards reviewed by Jonson et al. (2019).

Guidelines and case study. The M-KIT (Modularer Kurzintelligenztest; Dantlgraber et al., 2015) was developed to provide a time-efficient measure of cognitive ability while treating gender fairness as a construction principle rather than a post-hoc check. Its development illustrates the three decision levels at which fairness can be addressed—the theoretical model, the task format, and the item—and, critically, that these levels are hierarchically ordered: decisions taken at the higher levels constrain what remains achievable at the lower ones. This ordering is the organizing logic of our recommendations (Table 1), and we use the M-KIT here to make each level concrete. It is not offered as a novel instrument or as new evidence, but as one of the few documented cases in which fairness reasoning is visible at each level of test construction (Funsch and Roemer, 2025).

The theoretical model (Table 1, Recommendations 6–8). The theoretical model is the highest level of abstraction and the one with the greatest leverage over gender bias, because it fixes what the test is taken to measure and therefore what counts as construct-relevant in the first place. Because the validity of cognitive ability tests is attributed largely to the g factor rather than to specific abilities (Ree and Carretta, 2022), and because the criterion in applied settings is typically a broad construct such as overall job performance, the M-KIT targets fluid intelligence (g_f), defined in Cattell-Horn-Carroll (CHC) model terms as a content-general information-processing capacity that can in principle be applied to any content (Jensen, 1998; McGrew, 2009). This definition carries a direct fairness implication that is easily overlooked: If g_f is content-general, then content-specific variance is by definition not part of the construct, and any systematic group difference that is tied to a particular content format is therefore a candidate for construct-irrelevant variance rather than a true difference in the target ability. Yet figural tasks are routinely used as the sole estimate of g_f , even though verbal and numerical reasoning tasks show comparable or higher g-loadings (Johnson et al., 2004), and figural-only measurement yields larger gender differences than multi-content measurement (Hyde, 2005).

The M-KIT responds to this at the level of the model itself, through a modular design that operationalizes two distinct—and often conflated—definitions of fairness. The first core-focused

Table 1 A gendered innovations approach to fair test development: recommendations by stage, with graduated implementation.

#	Stage/ Recommendation	Rationale	Minimal implementation	Full implementation
A. Research priorities and assumptions (before the construct is defined)				
1	Relocate fairness work to the design stage	Fairness evidence is concentrated downstream, where the operationalization is already fixed: only 3 of 17 established fairness recommendations apply at the design stage (Jonson et al., 2019). This is not a generic “consider quality early” injunction: post-hoc item selection has been shown not to repair group gaps that originate in the operationalization (Bazian et al., 2025), so early action is necessary rather than merely advisable.	Document, before item writing, which construct facets and stimulus formats are being included and excluded, and why.	Treat construct definition as a formal fairness decision point with recorded justification, as in the M-KIT’s hierarchical fairness-decision model (Dantigraber et al., 2015).
2	Question default assumptions	A “default” is an option adopted when no alternative is specified (Schiebinger, 2014). Defaults about which stimuli are “neutral” or “efficient” are where gendered assumptions enter unexamined.	Ask explicitly why the intended format is standard, and what an alternative would look like.	Conduct a structured reflexivity step on the developer team’s own assumptions and positionality before construct definition (Schiebinger, 2014; Tannenbaum et al., 2019).
3	Adopt interdisciplinary frameworks	Cycle-spanning frameworks supply the interrogation of the construct definition that design-stage psychometric frameworks (e.g., Evidence-Centered Design; Mislevy & Haertel, 2006) do not, since the latter secure coverage of the construct as defined.	Use one framework (Gendered Innovations) as a checklist across development stages.	Combine with intersectional (Hankivsky et al., 2014) and community-engaged (Mohamed et al., 2026) frameworks to address multiple, intersecting fairness dimensions.
4	Incorporate diverse expertise in review	Homogeneous developer teams reproduce their own defaults; subgroup expertise raises sensitivity to biases they cannot see. Engagement is a continuum, not a binary (Mohamed et al., 2026).	Consult a review panel with subgroup expertise on a fixed draft of the construct and items.	Involve subgroup experts and affected stakeholders as co-developers at construct definition and format selection — the stages where decisions are still reversible.
5	Design for practical constraints	Recruitment settings demand short, inexpensive instruments; ignoring this guarantees that fair designs go unused, and resource pressure is precisely what drives narrow single-format tests.	State the time and cost budget explicitly and treat it as a design constraint rather than an afterthought.	Achieve breadth within the budget by selecting few, high-g-loading tasks across content domains, as in the M-KIT’s modular design (Dantigraber et al., 2015).
B. Construct definition and theoretical model				
6	Broaden construct representation	Narrow operationalization is construct underrepresentation—one of the two principal threats to validity (Messick, 1989, 1995). Where it is systematically patterned by group, it constitutes bias by design.	Sample the construct across at least verbal, numerical, and figural content rather than a single format (Jäger et al., 1997; Waschl and Burns, 2020).	Derive content coverage from an explicit theoretical model of the construct, justifying each facet’s inclusion by g-loading and construct relevance (Johnson et al., 2004; McGrew, 2009).
7	Treat fairness as a component of validity, not a trade-off	A test that systematically misrepresents a group’s standing on the construct is, to that extent, not valid for that interpretation (Kane, 2010; Messick, 1995). Fairness and validity are therefore not competing criteria.	Report fairness evidence within the validity argument rather than as a separate section.	Build the validity argument so that fairness evidence is constitutive of it, and revisit the construct definition when the two appear to conflict.
8	Use group differences as a diagnostic, not a target	On latent, content-general constructs for which there is independent warrant to expect no true group difference (e.g., fluid intelligence; Colom et al., 2000; Giofrè et al., 2022), a group gap tied to one content format signals construct-irrelevant method variance (Campbell and Fiske, 1959)—not something to be equalized for its own sake. The aim is removing construct-irrelevant variance; gap reduction is its observable signature.	Before treating a group difference as substantive, ask whether it is confined to one stimulus format.	Compare group differences across content formats loading on the same latent factor, and treat format-specific gaps as evidence of a design defect requiring re-operationalization.
C. Task and stimulus format				
9	Select formats that minimize construct-irrelevant demands	Figural matrices recruit visuo-spatial ability beyond the target construct, and the size of the gap depends on manipulable item-design features (Arendasy and Sommer, 2012; Waschl and Burns, 2020)—which a genuine trait difference would not.	Screen candidate formats for known construct-irrelevant demands (e.g., mental rotation).	Develop formats that reduce those demands while preserving g-loading, and balance formats showing opposing group advantages, as in the M-KIT’s two modules (Dantigraber et al., 2015).
D. Items, analysis, and interpretation				
10	Conduct routine bias checks — while recognizing their limits	DIF, measurement invariance, and predictive-bias analysis are necessary but operate on an already-fixed operationalization. A single-format test can be scalar-invariant across groups and still under-represent the construct; invariance and underrepresentation are orthogonal (Aguinis and Culpepper, 2024; Bauer, 2017; Leitgöb et al., 2023; Wetzel and Heli, 2013).	Run DIF and invariance analyses as standard practice and report them.	Run them after the design-stage decisions above, and interpret null results as evidence about the chosen items only — never as evidence that the construct was adequately defined.
11	Do not treat group-specific norms as a fairness remedy	Norms adjust the score distribution without touching the instrument. If the difference is construct-irrelevant variance, norming conceals a measurement artifact; if it is genuine, norming makes the outcome depend on group membership, which in selection contexts threatens equal treatment and is legally prohibited in some jurisdictions (Sackett and Wilk, 1994).	Report subgroup distributions transparently rather than norming them away.	Address the disparity at its source in the operationalization; reserve norms for interpretive purposes outside high-stakes selection.

Table 1 (continued)			
#	Stage/Recommendation	Rationale	Full implementation
E. Documentation and standards			
12	Document fairness transparently	Fairness decisions made at the design stage are invisible in the finished instrument unless recorded; most manuals report reliability and norms but little fairness reasoning (Jonson et al., 2019).	Document the full decision hierarchy—model, format, item—including alternatives considered and rejected (Dantlgraber et al., 2015).
13	Advance fairness within testing standards	Fairness is foundational in the current Standards (AERA et al., 2014; SIOP, 2018) but remains subordinated to objectivity in German-speaking standards (DIN, 2016; Bühner et al., 2017; Moosbrugger and Kelava, 2020), and design-stage procedures are underdeveloped in both.	Advocate for design-stage fairness procedures in standards revision, particularly in the German-speaking context.
Recommendations are ordered by stage of the research and development cycle; stages B–D correspond to the decision levels illustrated by the M-KIT (theoretical model, task format, item level; Dantlgraber et al., 2015). The graduated implementation columns follow the logic that engagement with fairness is a continuum constrained by resources rather than an all-or-nothing commitment (adapted from Mohamed et al., 2026). Recommendation 8 applies only to latent, content-general constructs for which there is independent theoretical and empirical warrant to expect no true group difference at the level of the core construct; constructs whose group differences are constitutive rather than artifactual (e.g., reaction time across age groups) fall outside its scope. Stakeholder and subgroup input (Recommendation 4) targets construct-irrelevant features of the operationalization—stimulus content and format—rather than the theoretical core of the construct, which remains disciplined by evidence on g-loading and construct relevance.			

module samples three high-g-loading tasks spanning verbal, numerical, and figural reasoning. Its rationale is validity-driven: by representing the construct across multiple content domains rather than a single format, it broadens construct representation so that the shared, content-general variance (g_c) is what the composite captures, while content-linked variance—including the visuo-spatial demands specific to figural material—is diluted rather than allowed to masquerade as ability. On this logic, any narrowing of the gender gap is the diagnostic signature that construct-irrelevant variance has been removed. The second module complements the first with three additional, lower-g-loading tasks selected so that formats showing a small male advantage are offset by formats showing a small female advantage or none, yielding a balanced overall score. This second configuration pursues fairness in the distinct sense of minimizing mean group differences directly, at the level of the composite. The two modules thus embody two different answers to the question “what would make this test fair?”—the elimination of construct-irrelevant variance versus the equalization of outcomes—and it is worth being explicit that these are not the same commitment.

The task format (Table 1, Recommendation 9). Where the model level fixes what is measured, the format level fixes how and thereby determines which construct-irrelevant demands a task imposes on the test taker. Performance differences on conventional figural matrices are partly attributable to mental rotation and visuo-spatial processing—capacities shaped by experience and socialization rather than by fluid reasoning itself (Kheloui et al., 2023; Niileksela et al., 2013)—so that a figural item can require of the test taker something over and above the inference the construct is meant to capture. The M-KIT retains figural stimuli, which are efficient and highly g-loaded, but redesigns the format to strip out that surplus demand: in place of systematically varied geometric figures that must be mentally manipulated (Raven et al., 1997), it uses easily perceived pictorial content, so that solving an item depends on sequential, rule-based inference rather than on spatial manipulation (cf. Hegarty, 2017). The construct is preserved, and the efficiency of the figural format is retained; the construct-irrelevant demand is designed out. This illustrates that broadening representation and refining formats are complementary rather than competing strategies – the format can be improved even within a content type that a purely model-level analysis might have discarded.

The item level (Table 1, Recommendation 10). Only after these higher-level decisions were fixed did Dantlgraber et al. (2015) turn to the item level, conducting DIF analyses to detect items that function differently across groups at equal trait levels (Osterlind and Everson, 2009) and excluding those that favored one gender. The item level is the least powerful of the three levers, precisely because the decisions taken at the model and format levels have already bounded what item selection can accomplish: DIF can remove individual biased items, but it cannot widen a construct that was defined too narrowly or repair a format that imposes a construct-irrelevant demand. Where such a format has already been chosen, removing individual items yields little or no improvement in fairness (Bazian et al., 2025). This is the same structural point that limits the analytic tools dominating current standards—they operate on a fixed operationalization—and it explains why, in the M-KIT, item-level screening functions as a final refinement of an already fair design rather than as the mechanism by which fairness is achieved.

Discussion

In this article, we argue that psychological tests require a fairness-centered approach from the outset—especially in organizational contexts—to prevent systematically disadvantaging certain groups

in personnel selection or development. While existing discussions often focus on reducing bias after testing, we emphasize a “bias by design” problem rooted in overly narrow conceptualizations of abilities and efficiency-driven choices that ignore the multifaceted nature of constructs (construct underrepresentation; Messick, 1989, 1995). The mechanism is not peculiar to figural matrices. It recurs wherever a discretionary feature of the operationalization, rather than the construct itself, determines the size of the group difference—and it can be located at each of the three decision levels described below. At the level of the theoretical model, for example, sex differences in declarative knowledge vary substantially with which knowledge domains are sampled, with women advantaged in health-related content and men in natural-science content (Steinmayr et al., 2015); the “general” in general knowledge conceals a content choice. At the level of the task format, item format alone explains roughly a quarter of the variation in gender achievement gaps across US states, with multiple-choice formats favouring men and constructed-response formats favoring women (Reardon et al., 2018), a pattern replicated across PISA cycles and 34 countries (Shear, 2023); and time limits inflate the male advantage in mental rotation, which is reduced—though not eliminated—when the constraint is removed (Maeda and Yoon, 2013). At the level of scoring, removing the penalty for incorrect answers from Chile’s national college entrance examination narrowed the gender gap in both skipped questions and scores, since penalties disadvantage the more risk-averse at equal levels of knowledge (Coffman and Klinowski, 2020)—a design feature that alters no item content at all. In each case, a defensible-looking default produced a group difference that a different, equally defensible default would not have produced. This is what we mean by bias by design, and it is why the remedy must be located where the default is chosen.

Building on the frameworks of procedural and distributive justice, we demonstrate that current methods for ensuring fairness are inadequate when they merely replicate the same standardized procedures without considering group-specific outcomes. Moreover, findings indicating that predictive validity estimates of cognitive ability tests can be lower than previously assumed (Sackett et al., 2024) underscore the need to treat fairness not as a trade-off for validity but as integral to it. Therefore, our approach includes recommendations for developing gender-sensitive tests, guided by the Gendered Innovations approach, to highlight how genuinely valid assessment tools must also be fair—enhancing both the ethical and practical dimensions of personnel selection. A recent study examined whether selecting specific items from existing cognitive ability tests—those with lower group mean differences—could reduce bias between ethnic subgroups (Bazian et al., 2025). However, the study found no convincing evidence supporting this approach, questioning its effectiveness in creating a fairer test after initial item construction and administration. This indicates the need for including gender and fairness aspects much earlier in psychological test development for selection processes than the stage of analysis. To integrate gender dimensions earlier in the research process, we referred to the Gendered Innovations approach as a social-constructive approach. Researchers are encouraged to comprehensively reflect upon their own presumptions when building the theoretical framework and test material in psychological testing.

An illustrative example of how the Gender Innovations approach can be applied to psychological testing is the M-KIT (Dantlgraber et al., 2015). Here, the authors provide a comprehensive assessment of gender fairness, rooting out the underlying theoretical assumptions. Enhancing fairness considerations beyond the test administration and item formatting can thus provide novel theoretical insights and strengthen the psychometric quality of developed tests. As outlined above, guidelines

for reporting and administering tests are already quite elaborate (Jonson et al., 2019). Thus, we specifically focus on the fundamental assumptions that guide developers and integrate practices of the Gendered Innovations approach.

Introducing best practices for integrating fairness perspectives in cognitive ability tests. Even though the measure by Dantlgraber et al. (2015) set an important precedent for gender-sensitive cognitive ability tests over a decade ago, to our knowledge, no cognitive ability test developed with gender fairness as an explicit design principle has followed in the years since, highlighting both the continued relevance of this instrument and the urgent need for further test development efforts in this space. The model and task format level in their fairness approach rethinks established theories of intelligence, thus enabling more nuanced assessment of gender fairness. Past guidelines overtly address the item level of the final developed tests. Changes at this stage may have less influence on gender fairness, as the greater bias originates from the theoretical foundation of tests.

As shown by Jonson et al. (2019) and also discussed by Dantlgraber et al. (2015), past tests rarely provide compelling evidence when it comes to fairness aspects in their test construction and administration. Often, established instruments rely on reliability assessments or refer to separate norms for men and women, lacking validity indices or more elaborate analysis steps as described above. When practitioners rely on norm values for men and women, problems arise as the test performance and decisions of individuals (e.g., applicants) are no longer independent of the applicant’s gender, threatening equal treatment of men and women. Thus, the M-KIT offers a more holistic approach. The underlying model rethinks concepts of fairness in theoretical foundations of test development, starting not at the developed product but at the underlying concepts and assumptions of intelligence, its assessment, and gender fairness. In Table 1, we derive best practices based on the above elaborations on integrating fairness perspectives in psychological test development for applied settings.

Utilizing these recommendations to rethink concepts and theories on a theoretical level (which cognitive abilities do we want to assess with which test material?) and on a methodological level (which quality criteria guide our research designs?) enables a more elaborate evaluation of fairness in psychological test development. These initial steps influence the hypotheses, methodology, test material, data collection, data analysis, and finally the test scores derived in application settings and decisions on human biographies.

Limitations. Our argument is nonetheless subject to several limitations. Wherever a construct is defined and operationalized, design-stage assumptions can encode a gendered default. While the presented guidelines emphasize fairness in cognitive ability testing, it is important to acknowledge the domain-specific nature of our example. Cognitive ability tests, especially those assessing fluid intelligence, allow for a relatively high degree of freedom in item and task design. This flexibility facilitates the development of theoretically grounded and fairness-sensitive assessments. However, such flexibility may not be equally available in other psychological measurement domains. For instance, in vocational interest inventories, attempts to optimize gender fairness via item-level DIF reduction can lead to unintended shifts in construct validity (Wetzel and Hell, 2013). For example, the “realistic” interest domain of the RIASEC model (Holland, 1997) predominantly features items related to manual labor and the use of tools and machines, which are typically endorsed by men. To better reflect the experiences of both men and women, this

domain should include a broader spectrum of activities such as gardening or cooking and additionally incorporate tasks that utilize tools and machines to enhance healthcare or address environmental issues (Einarsdottir and Rounds, 2009). However, this approach would alter the established definition of “realistic” interests. This raises questions about whether fairness-driven item adjustments might compromise construct coherence, suggesting a need for domain-specific fairness strategies that are sensitive to construct integrity. Additionally, this example as well represents a potentially biased conceptualization of the “realistic” interest domain and strengthens our case for considering fairness on the level of theory and constructs. Situational judgement tests are instructive in a different way: gender differences are small and, where present, tend to favour women, whereas ethnic subgroup differences grow with the cognitive loading of the test (Bardach et al., 2021; Herde et al., 2020; Whetzel et al., 2008). That moderation is itself evidence for our argument, since the magnitude of the group difference tracks how the construct is operationalized rather than the attribute it is meant to capture. Our imperative to address fairness at the earliest stage of instrument development therefore extends beyond cognitive ability testing, even though we develop it here for that case.

A second limitation concerns the treatment of gender within our framework. Throughout this article, gender is operationalized as a binary variable (men/women), which reflects common practice in psychological testing and the current availability of normative data. However, this binary view excludes individuals identifying as nonbinary, gender-diverse, or “divers” (as categorized in some German data collections). Given the increasing recognition of gender diversity, future test development and validation processes must account for these identities to ensure inclusive and fair assessment practices (Kheloui et al., 2023). This presents methodological and practical challenges, particularly due to the underrepresentation of these groups in most empirical samples, which can hinder robust subgroup analyses and DIF testing. Addressing this limitation requires both theoretical refinement of gender constructs and intentional oversampling strategies in test standardization efforts.

A further limitation concerns our reliance on a single, established instrument as the primary illustration. The M-KIT was published in 2015, and its construction principles were documented at that time; we therefore do not offer it as a novel instrument or as new empirical evidence, and structuring the argument around one example inevitably narrows the demonstration and risks reading as advocacy for a particular test. We use it instead as a worked example of design-stage fairness reasoning—a rare case in practice and, to our knowledge, among the first cognitive ability manuals in the German-speaking context to include an explicit gender-fairness chapter and a hierarchical fairness-decision model. The intended contribution is not the instrument, but the transferable design principles abstracted from it through the Gendered Innovations framework (Table 1), which this example makes concrete without exhausting. Future work should demonstrate these principles across additional instruments and constructs, both to establish their generality and to reduce dependence on any single exemplar.

Conclusion

In sum, we contend that truly fair psychological tests—especially those used for high-stakes decisions in education and personnel selection—cannot be achieved solely through standardized administration or post-hoc statistical assessments or adjustments. Instead, fairness must be built into every layer of test development, from the theoretical foundations that define what we measure, to the task formats that determine how we measure it, down to the individual items that compose each scale. Drawing on the Gendered

Innovations framework and the case of the M-KIT, we have demonstrated how a constructivist, gender-sensitive approach can broaden construct representation, question default assumptions, and balance both procedural and distributive fairness. By integrating gender considerations early—rethinking models of fluid intelligence, diversifying content domains, and rigorously screening for DIF—test developers can create more excellent instruments that not only predict performance but do so equitably across all groups.

Looking forward, we urge test-makers and standard-setting bodies to adopt comprehensive, interdisciplinary guidelines that elevate fairness to the same status as reliability and validity. Future research should extend this approach to nonbinary and intersectional identities, as well as to psychological test domains where theoretical flexibility is more constrained. Ultimately, fairness is not a trade-off against validity; it is an essential facet of it. Only by embedding fairness at the very heart of test theory, design, and documentation can we ensure that psychological assessments serve as tools of opportunity rather than instruments of inadvertent exclusion.

Data availability

No datasets were generated or analysed during the current study.

Received: 23 July 2025; Accepted: 7 September 2026;

Published online: 25 September 2026

Notes

- 1 Our constructivist framing concerns the choice of operationalization – which construct facets are represented, and through which stimuli – not the measurement model applied thereafter, and it does not deny that individual differences in fluid intelligence are real. The point in fact follows from classical test theory: the true score is defined as the expected observed score over replications of a *given* test (Lord and Novick, 1968) and is therefore procedure-relative rather than instrument-independent; equating it with the construct is a recognized error (Borsboom and Mellenbergh, 2002). A true score on a figural-only test corresponds to expected figural-reasoning performance, not fluid intelligence.
- 2 Other forms of justice, such as informational and interpersonal justice (Colquitt et al., 2001), have been identified within the perceived justice literature. In this paper, we chose to only focus on procedural and distributive justice as other forms of justice are not relevant to the issues discussed here (Kane, 2010).

References

- Aguinis H, Culpepper SA (2024) Improving our understanding of predictive bias in testing. *J Appl Psychol* 109(3):402–414. <https://doi.org/10.1037/apl0001152>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014) Standards for educational and psychological testing. Washington, D.C.: American Psychological Association
- Arendasy ME, Sommer M (2012) Gender differences in figural matrices: the moderating role of item design features. *Intelligence* 40(6):584–597. <https://doi.org/10.1016/j.intell.2012.08.003>
- Armonit C, Schuler H, Hell B (2020) NutzuNg, Validität, Praktikabilität und Akzeptanz psychologischer Personalauswahlverfahren in Deutschland 1985, 1993, 2007, 2020. *Z Arb Organ AO* 64(2):67–82. <https://doi.org/10.1026/0932-4089/a000311>
- Azen R, Bronner S, Gafni N (2002) Examination of gender bias in university admissions. *Appl Meas Educ* 15(1):75–94. https://doi.org/10.1207/S15324818AME1501_05
- Bardach L, Rushby JV, Klassen RM (2021) The selection gap in teacher education: adverse effects of ethnicity, gender, and socio-economic status on situational judgement test performance. *Br J Educ Psychol* 91(3):e12405. <https://doi.org/10.1111/bjep.12405>
- Bauer DJ (2017) A more general model for testing measurement invariance and differential item functioning. *Psychol Methods* 22(3):507–526. <https://doi.org/10.1037/met0000077>
- Bazian IM, Lee SD, Sackett PR, Kuncel NR, Jacobs RR, McDaniel MA (2025) Prospects for reducing group mean differences on cognitive tests via item selection strategies. *J Appl Psychol* 110(6):876–886. <https://doi.org/10.1037/apl0001253>

- Berry CM, Lievens F, Zhang C, Sackett PR (2024) Insights from an updated personnel selection meta-analytic matrix: revisiting general mental ability tests' role in the validity-diversity trade-off. *J Appl Psychol* 109(10):1611–1634. <https://doi.org/10.1037/apl0001203>
- Borsboom D (2006) The attack of the psychometricians. *Psychometrika* 71(3):425–440. <https://doi.org/10.1007/s11336-006-1447-6>
- Borsboom D, Mellenbergh GJ (2002) True scores, latent variables, and constructs: a comment on Schmidt and Hunter. *Intelligence* 30(6):505–514. [https://doi.org/10.1016/S0160-2896\(02\)00082-X](https://doi.org/10.1016/S0160-2896(02)00082-X)
- Bühner M, Ziegler M, Kersting M (2017) Statistisch-methodische Grundlagen der Eignungsbeurteilung. In: Diagnostik- und Testkuratorium (Hrsg.), Personalauswahl kompetent gestalten (S. 155–188). 10.1007/978-3-662-53772-5_5
- Cabrera MAM, Nguyen NT (2001) Situational judgment tests: a review of practice and constructs assessed. *Int J Sel Assess* 9(1–2):103–113. <https://doi.org/10.1111/1468-2389.00167>
- Campbell DT, Fiske DW (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 56(2):81–105. <https://doi.org/10.1037/h0046016>
- Carroll JB (1993) Human cognitive abilities: a survey of factor-analytic studies (No. 1). Cambridge University Press
- Chung-Yan GA, Cronshaw SF (2002) A critical re-examination and analysis of cognitive ability tests using the Thorndike model of fairness. *J Occup Psychol* 75(4):489–509. <https://doi.org/10.1348/096317902321119709>
- Coffman KB, Klinowski D (2020) The impact of penalties for wrong answers on the gender gap in test scores. *Proc Natl Acad Sci USA* 117(16):8794–8803. <https://doi.org/10.1073/pnas.1920945117>
- Cole NS, Zieky MJ (2001) The new faces of fairness. *J Educ Meas* 38(4):369–382. <https://doi.org/10.1111/j.1745-3984.2001.tb01132.x>
- Colom R, Juan-Espinoso M, Abad F, García LF (2000) Negligible sex differences in general intelligence. *Intelligence* 28(1):57–68. [https://doi.org/10.1016/S0160-2896\(99\)00035-5](https://doi.org/10.1016/S0160-2896(99)00035-5)
- Colquitt JA (2001) On the dimensionality of organizational justice: a construct validation of a measure. *J Appl Psychol* 86(3):386
- Dantlgraber M, Hell B, Fischer F, Schult J (2015) M-KIT: Modularer Kurzintelligenztest. Hogrefe
- Der G, Deary IJ (2009) Age and sex differences in reaction time in adulthood: Results from the United Kingdom Health and Lifestyle Survey: correction. *Psychol Aging* 24(1):229–229. <https://doi.org/10.1037/a0015515>
- Deutsches Institut für Normung (2016) DIN 33430: Anforderungen an berufsbezogene Eignungsdiagnostik. Beuth Verlag
- Einarsdóttir S, Rounds J (2009) Gender bias and construct validity in vocational interest measurement: Differential item functioning in the Strong Interest Inventory. *J Vocat Behav* 74(3):295–307. <https://doi.org/10.1016/j.jvb.2009.01.003>
- Engineering Innovation Processes | Gendered Innovations. (o. D.). <https://genderedinnovations.stanford.edu/methods/innovation.html>
- Farr JL, Tippins NT (2017) Handbook of Employee Selection. Taylor & Francis
- Flynn JR, Rossi-Casé L (2011) Modern women match men on Raven's Progressive Matrices. *Pers Individ Differ* 50(6):799–803. <https://doi.org/10.1016/j.paid.2010.12.035>
- Formann AK, Piswanger K (1979) Wiener Matrizen-Test (WMT). Beltz Test
- Funsch K, Roemer L (2025) TBS-DTK-rezension: "Modularer Kurzintelligenztest (M-KIT)". *Psychol Rundsch* 76(1):100–102. <https://doi.org/10.1026/0033-3042/a000701>
- Gignac GE (2015) Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of g. *Intelligence* 52:71–79. <https://doi.org/10.1016/j.intell.2015.07.006>
- Gilliland SW (1994) Effects of procedural and distributive justice on reactions to a selection system. *J Appl Psychol* 79(5):691–701. <https://doi.org/10.1037/0021-9010.79.5.691>
- Giofrè D, Allen K, Toffalini E, Caviola S (2022) The impasse on gender differences in intelligence: a meta-analysis on WISC batteries. *Educ Psychol Rev* 34(4):2543–2568. <https://doi.org/10.1007/s10648-022-09705-1>
- Greenberg J (1990) Organizational justice: yesterday, today, and tomorrow. *J Manag* 16(2):399–432. <https://doi.org/10.1177/014920639001600208>
- Greenberg J, Tyler TR (1987) Why procedural justice in organizations?. *Soc Justice Res* 1(2):127–142. <https://doi.org/10.1007/BF01048012>
- Hankivsky O, Grace D, Hunting G, Giesbrecht M, Fridkin A, Rudrum S, Ferlatte O, Clark N (2014) An intersectionality-based policy analysis framework: critical reflections on a methodology for advancing equity. *Int J Equity Health* 13(1). <https://doi.org/10.1186/s12939-014-0119-x>
- Hegarty M (2017) Ability and sex differences in spatial thinking: What does the mental rotation test really measure?. *Psychon Bull Rev* 25(3):1212–1219. <https://doi.org/10.3758/s13423-017-1347-z>
- Herde CN, Lievens F, Jackson DJR, Shalrooshan A, Roth PL (2020) Subgroup differences in situational judgment test scores: evidence from large applicant samples. *Int J Sel Assess* 28(1):45–54. <https://doi.org/10.1111/ijss.12269>
- Herzberg PY (2024) A critical examination of the claim that women are systematically disadvantaged. *Z Arb Organ AO*. <https://econtent.hogrefe.com/doi/10.1026/0932-4089/a000438>
- Holland JL (1997) Making vocational choices: a theory of vocational personalities and work environments (3rd ed.). Prentice-Hall
- Hossiep R, Schecke J, Weiß S (2015) Zum Einsatz von persönlichkeitsorientierten Fragebogen. Eine Erhebung unter den 580 größten deutschen Unternehmen. *Psychol Rundsch* 66(2):127–129
- Hossiep R, Turck D, Hasella M (1999) Bochumer Matrizen-test (BOMAT) Advanced. Hogrefe
- Hough LM, Oswald FL, Ployhart RE (2001) Determinants, detection and amelioration of adverse impact in personnel selection procedures: issues, evidence and lessons learned. *Int J Sel Assess* 9(1–2):152–194. <https://doi.org/10.1111/1468-2389.00171>
- Höft S, Püttner I, Kersting M (2017) Anforderungsanalyse, Verfahren der Eignungsbeurteilung sowie rechtliche Rahmenbedingungen. In: Diagnostik- und Testkuratorium (Hrsg.), Personalauswahl kompetent gestalten (S. 95–154). https://doi.org/10.1007/978-3-662-53772-5_5
- Hyde JS (2005) The gender similarities hypothesis. *Am Psychol* 60(6):581–592. <https://doi.org/10.1037/0003-066x.60.6.581>
- Jäger AO, Süß HM, Beauducel A (1997) Berliner Intelligenzstruktur-Test: BIS-Test
- Jensen AR (1998) The g factor: the science of mental ability. Human evolution, behavior, and intelligence. Praeger
- Johnson W, Bouchard TJ, Krueger RF, McGue M, Gottesman II (2004) Just one g: consistent results from three test batteries. *Intelligence* 32(1):95–107. [https://doi.org/10.1016/S0160-2896\(03\)00062-x](https://doi.org/10.1016/S0160-2896(03)00062-x)
- Jonson JL, Trantham P, Usher-Tate BJ (2019) An evaluative framework for reviewing fairness standards and practices in educational tests. *Educ Meas Issues Pract* 38(3):6–19. <https://doi.org/10.1111/emip.12259>
- Kane M (2010) Validity and fairness. *Lang Test* 27(2):177–182. <https://doi.org/10.1177/0265532209349467>
- Keller EF (1995) Reflections on gender and science. Yale University Press
- Kheloui S, Jacmin-Park S, Larocque O, Kerr P, Rossi M, Cartier L, Juster R (2023) Sex/gender differences in cognitive abilities. *Neurosci Biobehav Rev* 152:105333. <https://doi.org/10.1016/j.neubiorev.2023.105333>
- Koch M, Spinath FM, Greiff S, Becker N (2022) Development and validation of the open matrices item bank. *J Intell* 10(3):41. <https://doi.org/10.3390/jintelligence10030041>
- Kulikowski K, Cucina JM, Hayes TL, Oh I-S, Ones D, Viswesvaran C (2025) In defence of cognitive ability testing: affirming the evidence for its continued use in personnel selection and admission decisions. *J Occup Organ Psychol* 98(3):e70048. <https://doi.org/10.1111/joop.70048>
- Leitgöb H, Seddig D, Asparouhov T, Behr D, Davidov E, De Roover K, Jak S, Meitinger K, Menold N, Muthén B, Rudnev M, Schmidt P, van de Schoot R (2023) Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Soc Sci Res* 110:102805. <https://doi.org/10.1016/j.ssresearch.2022.102805>
- Lord FM, Novick MR, Birnbaum A (1968) Statistical theories of mental test scores. Addison-Wesley
- Lynn R, Irwing P (2004) Sex differences on the progressive matrices: a meta-analysis. *Intelligence* 32(5):481–498. <https://doi.org/10.1016/j.intell.2004.06.008>
- Maeda Y, Yoon SY (2013) A meta-analysis on gender differences in mental rotation ability measured by the Purdue Spatial Visualization Tests: visualization of rotations (PSVT:R). *Educ Psychol Rev* 25(1):69–94. <https://doi.org/10.1007/s10648-012-9215-x>
- McGrew KS (2009) CHC theory and the human cognitive abilities project: standing on the shoulders of the giants of psychometric intelligence research. *Intelligence* 37(1):1–10. <https://doi.org/10.1016/j.intell.2008.08.004>
- Messick S (1989) Meaning and values in test validation: the science and ethics of assessment. *Educ Res* 18(2):5–11. <https://doi.org/10.3102/0013189X018002005>
- Messick S (1995) Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 50(9):741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Mislevy RJ, Haertel GD (2006) Implications of evidence-centered design for educational testing. *Educ Meas Issues Pract* 25(4):6–20. <https://doi.org/10.1111/j.1745-3992.2006.00075.x>
- Mohamed J, Jacob BK, Debrosse R, Adel M, Zhang S, Hu Z, Harris K, Chung JM (2026) A practical guide for integrating community-engaged research across the psychological research cycle. *Adv Methods Pract Psychol Sci* 9(1):25152459251397997. <https://doi.org/10.1177/25152459251397997>
- Moosbrugger H, Kelava A (2020) Qualitätsanforderungen an Tests und Fragebogen („Gütekriterien“). In: Moosbrugger H, Kelava A (eds) Testtheorie und Fragebogenkonstruktion. Springer, pp 13–38. https://doi.org/10.1007/978-3-662-61532-4_2
- Murphy K (2017) What can we learn from “Not much more than g”? *J Intell* 5(1):8. <https://doi.org/10.3390/jintelligence5010008>

- Niileksela CR, Reynolds MR, Kaufman AS (2013) An alternative Cattell–Horn–Carroll (CHC) factor structure of the WAIS-IV: age invariance of an alternative model for ages 70–90. *Psychol Assess* 25(2):391–404. <https://doi.org/10.1037/a0031175>
- Osterlind S, Everson H (2009) Differential item functioning (2nd ed.). In: SAGE Publications, Inc. eBooks. <https://doi.org/10.4135/9781412993913>
- Plake BS, Wise LL (2014) What is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing?. *Educ Meas Issues Pract* 33(4):4–12. <https://doi.org/10.1111/emip.12045>
- Ployhart RE, Holtz BC (2008) The diversity–validity dilemma: strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Pers Psychol* 61(1):153–172. <https://doi.org/10.1111/j.1744-6570.2008.00109.x>
- Pollard E, Hirsh W, Williams M, Buzzeo J, Marvell RA, Tassinari A, Bertram C, Fletcher L, Artess J, Redman J, Ball C (2015) Understanding employers' graduate recruitment and selection practices. Advisory Conciliation and Arbitration Service. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/474251/BIS-15-464-employer-graduate-recruitment.pdf
- Prochazka J, Novacek J, Vaculik M (2023) The evaluation of personnel selection methods by HR practitioners: the effect of reference and its interaction with information about validity. *Int J Hum Resour Manag* 34(16):3233–3251. <https://doi.org/10.1080/09585192.2022.2117563>
- Pyburn JrKM, Ployhart RE, Kravitz DA (2008) The diversity–validity Dilemma: overview and legal context. *Pers Psychol* 61(1):143–151. <https://doi.org/10.1111/j.1744-6570.2008.00108.x>
- Raven JC (1939) The R.E.C.I. series of perceptual tests: an experimental survey. *Br J Med Psychol* 18(1):16–34. <https://doi.org/10.1111/j.2044-8341.1939.tb00705.x>
- Raven JC, Raven J, Court, JH (1997) Raven's progressive matrices and vocabulary scales. J.C. Raven Ltd
- Reardon SF, Kalogrides D, Fahle EM, Podolsky A, Zárate RC (2018) The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades. *Educ Res* 47(5):284–294. <https://doi.org/10.3102/0013189X18762105>
- Ree MJ, Carretta TR (2022) Thirty years of research on general and specific abilities: still not much more than g. *Intelligence* 91:101617. <https://doi.org/10.1016/j.intell.2021.101617>
- Reynolds CR, Suzuki LA (2012) Bias in psychological assessment. In: *Handbook of psychology*, second edition. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118133880.hop210004>
- Reynolds CR, Altmann RA, Allen DN (2021) The problem of bias in psychological assessment. In: Reynolds CR, Altmann RA, Allen DN (eds) *Mastering modern psychological testing: theory and methods*. Springer International Publishing, pp 573–613 https://doi.org/10.1007/978-3-030-59455-8_15
- Rohde TE, Thompson LA (2007) Predicting academic achievement with cognitive ability. *Intelligence* 35(1):83–92. <https://doi.org/10.1016/j.intell.2006.05.004>
- Sackett PR, Demeke S, Bazian IM, Griebie AM, Priest R, Kuncel NR (2024) A contemporary look at the relationship between general cognitive ability and job performance. *J Appl Psychol* 109(5):687–713. <https://doi.org/10.1037/apl0001159>
- Sackett PR, Wilk SL (1994) Within-group norming and other forms of score adjustment in preemployment testing. *Am Psychol* 49(11):929–954. <https://doi.org/10.1037/0003-066x.49.11.929>
- Sackett PR, Zhang C, Berry CM, Lievens F (2023) Revisiting the design of selection systems in light of new findings regarding the validity of widely used predictors. *Ind Organ Psychol* 16(3):283–300. <https://doi.org/10.1017/iop.2023.24>
- Schiebinger L, Klinge I, Paik HY, Sánchez de Madariaga I, Schraudner M, and Stefanick M (eds) (2011–2020) *Gendered innovations in science, health & medicine, engineering, and environment* (genderedinnovations.stanford.edu)
- Schiebinger L (2014) Scientific research must take gender into account. *Nature* 507(7490):9. <https://doi.org/10.1038/507009a>
- Schmidt FL, Hunter JE (1974) Racial and ethnic bias in psychological tests: divergent implications of two definitions of test bias. *Am Psychol* 29(1):1–8. <https://doi.org/10.1037/h0035844>
- Schmidt FL, Hunter JE (1998) The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. *Psychol Bull* 124(2):262–274. <https://doi.org/10.1037/0033-2909.124.2.262>
- Schmidt-Atzert L (2004) *Objektiver Leistungsmotivations-Test OLMT [Objective Achievement-Motivation-Test OLMT]*. Mödling: Schuhfried
- Shear BR (2023) Gender bias in test item formats: evidence from PISA 2009, 2012, and 2015 math and reading tests. *J Educ Meas* 60(4):676–696. <https://doi.org/10.1111/jedm.12372>
- Society for Industrial and Organizational Psychology (2018) Principles for the validation and use of personnel selection procedures. *Ind Organ Psychol* 11(S1):1–97. <https://doi.org/10.1017/iop.2018.195>
- Sparfeldt JR, Schult J (2024) Mean differences in test scores, gender bias, and threats to fairness. *Z Arb Organ AO*. <https://doi.org/10.1026/0932-4089/a000439>
- Speer AB, Hickman L, Song QC, Perrotta J, Jacobs RR, Lambert D (2025) Addressing the diversity–validity dilemma in personnel selection: unraveling the impact of multipenalty optimized regression in varied testing scenarios. *J Appl Psychol*. <https://doi.org/10.1037/apl0001282>
- Steel P, Fariborzi H (2024) Supplemental material for a longitudinal meta-analysis of range restriction estimates and general mental ability validity coefficients: better addressing overcorrection amid decline effects. *J Appl Psychol* 109(12):1901–1920. <https://doi.org/10.1037/apl0001214.supp>
- Steinmayr R, Bergold S, Margraf-Stiksrud J, Freund PA (2015) Gender differences on general knowledge tests: Are they due to Differential Item Functioning?. *Intelligence* 50:164–174. <https://doi.org/10.1016/j.intell.2015.04.001>
- Striebing C, Schneider J, Uhler L, Hartmann F (2024) The Fairness Imperative. *Z Arb Organ AO*. <https://doi.org/10.1026/0932-4089/a000437>
- Stumpf S, Leenen WR, Scheitza A (2017) Adverse impact in der Personalauswahl einer deutschen Behörde: eine Analyse ethnischer Subgruppendifferenzen. *Ger J Hum Resour Manag* 31(1):4–31. <https://doi.org/10.1177/2397002216637289>
- Tannenbaum C, Ellis RP, Eyssel F, Zou J, Schiebinger L (2019) Sex and gender analysis improves science and engineering. *Nature* 575(7781):137–146. <https://doi.org/10.1038/s41586-019-1657-6>
- Waschl N, Burns NR (2020) Sex differences in inductive reasoning: a research synthesis using meta-analytic techniques. *Personal Individ Differ* 164:109959. <https://doi.org/10.1016/j.paid.2020.109959>
- Weiß RH (2019) CFT 20-R mit WS: Grundintelligenztest Skala 2-revision (CFT 20-R) mit Wortschatztest und zahlenfolgentest-revision. Hogrefe
- Wetzel E, Hell B (2013) Gender-related differential item functioning in vocational interest measurement. *J Individ Differ* 34(3):170–183. <https://doi.org/10.1027/1614-0001/a000112>
- Whetzel DL, McDaniel MA, Nguyen NT (2008) Subgroup differences in situational judgment test performance: a meta-analysis. *Hum Perform* 21(3):291–309. <https://doi.org/10.1080/08959280802137820>
- Wilson CJ, Bowden SC, Byrne LK, Joshua NR, Marx W, Weiss LG (2023) The cross-cultural generalizability of cognitive ability measures: a systematic literature review. *Intelligence* 98: 101751. <https://doi.org/10.1016/j.intell.2023.101751>
- Woods SA, Patterson F (2024) A critical review of the use of cognitive ability testing for selection into graduate and higher professional occupations. *J Occup Organ Psychol* 97(1):253–272. <https://doi.org/10.1111/joop.12470>
- Xi X (2010) How do we go about investigating test fairness?. *Lang Test* 27(2):147–170. <https://doi.org/10.1177/0265532209349465>

Acknowledgements

No funding was received for this research project.

Author contributions

J.S., C.S., K.P. and M.J. wrote the main manuscript text. J.S., C.S. and M.J. conceptualized the argumentative approach of the manuscript. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Declaration

Competing interests

Clemens Striebing is a Special Issue Guest Editor for this journal at the time of submission.

Ethics approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Jannick Schneider or Melanie Elizabeth Jacobsen.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026