Preliminary Program 19th Annual Meeting of the Dutch-Flemish Network for Selection Research

Friday, October 17th, 2025
Room: HG-0C29 Aurora, Main building, VU Amsterdam

09:30h – 09:50h	Arrival, coffee & tea
09:50h – 10:00h	Opening Rob Meijer & Reinout de Vries
10:00h – 11:00h	**Keynote:** Faking in Personality Tests and Other Selection Procedures: What Can We Do to Prevent It? (prof. dr. Klaus Melchers, Ulm University)
11:00h – 11:15h	Coffee break
11:15h – 11:40h	Perspectives of Students and Assessors of Standardization in Performance-based Assessments (Juan Arboleda)
11:40h – 12:05h	Performance-based Assessments in Higher Education: Stakeholders' Fairness Perceptions of Structured versus Unstructured Assessment Methods (Sanjay van Buel)
12:05h – 13:00h	**Lunch at 'Grand Café LIVING Amsterdam** (on Campus)
13:00h – 13:25h	Does Feedback on the Accuracy of Human and Algorithmic Prediction Increase Algorithm Use and Reduce Better-Than-Average-Effects Over Time? (Jacob Matić)
13:25h – 13:50h	Selection System Design: Exploring Predictor Choice and Predictor Importance for Validity and Diversity (Susan Niessen)
13:50h – 14:15h	The First Comprehensive Review of Resume Screening (Isolde van der Schuur)
14:15h – 14:40h	The Implicit Trait Function (Stijn Schelfhout)
14:40h – 15:00h	Coffee break
15:00h – 15:25h	Identity in Reflection: Studying How Gender and Minority Status Shape Medical Students' Narrative Self-Evaluations in Comparison to Faculty Evaluations (Marjolijn Wijnen)
15:25h – 15:50h	Creating and Testing Stimulus Materials: A Comparison between Text- and Video-based Stimuli (Aabidien Hassankhan)
15:50h – 16:35h	An Interview with the Upcoming and Outgoing Organizers of the Dutch-Flemish Meeting on Selection Research (Janneke Oostrom, Susan Niessen, Rob Meijer, & Marise Born)
16:35h – 16:45h	Closing
16:45h – 17:45h	Drinks at **Grand Café LIVING Amsterdam** (on Campus)
18:00h – 21:00h	**Dinner at Gustavino Restaurant & Vinoteca**, Gustav Mahlerplein 16, 1082 MA Amsterdam

Keynote: Faking in Personality Tests and Other Selection Procedures: What Can We Do to Prevent It? (10:00-11:00)

Klaus Melchers
Ulm University, Germany

Applicants usually want to make a good impression in selection situations to increase their chances for a job offer. To do so, they try to demonstrate actual skills and qualifications but, in many situations, applicants also resort to deceptive strategies and use faking in selection situations. Faking does not only impair the measurement of applicants' true skills and qualifications but is also related to negative personality characteristics such as dark personality traits and low honesty-humility. Accordingly, many researchers as well as organizations and selection practitioners are concerned that faking impairs the quality of selection decisions and this is not only the case for personality testing, where faking has mainly been discussed and investigated, but also for all other selection procedures in all of which faking and deceptive behavior can occur as well. Furthermore, the detection of faking is challenging and often hardly possible. Therefore, a relevant question for applied settings is whether faking can be prevented or at least be reduced in the first place. To provide an answer to this question, I will turn to common faking models, which—at their core—assume that faking depends on applicants' motivation, ability, and opportunity to fake. In my presentation, I will give an overview of research on different suggestions aiming to prevent faking by either trying to lower applicants' motivation to fake or their opportunity to fake. I will review evidence for the effectiveness of the different approaches to prevent or at least to reduce faking. Furthermore, I will also consider findings concerning other effects of the different approaches that are relevant for applied settings such as effects on psychometric properties or applicant reactions. I will conclude my presentation with a set of recommendations for practice and with recommendations for future research.

Perspectives of Students and Assessors of Standardization in Performance-based Assessments (11:15-11:40)

Juan Arboleda[1], Rob R. Meijer[1], Marvin Neumann[2], A. Susan M. Niessen[1]

[1]University of Groningen

[2]VU Amsterdam

Judgment standardization increases the reliability of performance-based assessment but often leads to resistance among stakeholders. In the educational domain, little is known how different components of judgment standardization are perceived by assessors and students. Therefore, in the present pre-registered study, we investigated the effects of various components of judgment standardization on both assessor and student perceptions. We applied a mixed-methods approach, using a quantitative between-subjects design (assessors *n* = 417, students *n* = 400) and qualitative focus groups (*n* = 20 assessors, *n* = 20 students). The qualitative data helped to understand reasons for perceptions and to identify themes beyond the perception constructs included in the quantitative part. We found the most positive perceptions (for both students and assessors) for (1) methods with high rating standardization (predefined criteria with definitions and anchored scales) as opposed to low rating standardization (only broadly defined criteria) and for (2) absolute judgment as compared to comparative judgment. Contrary to previous research, our findings suggest that low rating standardization may lead to more resistance for both students and assessors than high rating standardization. These results show the value of looking at distinct aspects of standardization, instead of the common dichotomy of holistic versus analytic judgment.

Performance-based Assessments in Higher Education: Stakeholders' Fairness Perceptions of Structured versus Unstructured Assessment Methods (11:40-12:05)

Sanjay W. L. van Buel[1], Karen M. Stegers-Jager[2], Janneke K. Oostrom[3], & Marise Ph. Born[1]

[1]Erasmus University Rotterdam

[2]Radboud University Nijmegen

[3]Tilburg University

Assessments are fundamental to student learning. Performance-based assessments (PBAs, such as literature reviews) are used to simultaneously evaluate key professional skills, such as analytical thinking and project management. However, a lack of consistent evaluation criteria in PBAs has led to inconsistent grading practices. A proposed solution is to better structure assessment methods, curbing the potential influence of irrelevant biasing factors on student outcomes. Students, especially ethnic minorities, report preferring structured over unstructured methods, while assessors report preferring more leeway in their assessment tools. This mismatch in perceptions, however, could negatively impact student evaluation and the implementation of such methods in practice. We therefore examined how structuring PBA-related assessment methods impacted stakeholders' perceptions of several fairness principles, and how students' minority status affected these perceptions. In a vignette study, data on 387 university students and 469 assessors were collected via an online survey. Corroborating previous studies, we found that both groups preferred more structured methods over less structured methods. While students' ethnic background did not influence assessors' perceptions, ethnic minority students reported consistently lower fairness ratings across vignettes compared to ethnic majority students. The above illustrates that structuring methods could be a first step to improve stakeholders' perceptions of such methods.

Does Feedback on the Accuracy of Human and Algorithmic Prediction Increase Algorithm Use and Reduce Better-Than-Average-Effects Over Time? (13:00-13:25)

Jacob J. Matić, Marvin Neumann, & Reinout E. de Vries
VU Amsterdam

The aim of this project is to investigate whether feedback on participants' predictive validity in a prediction task can stimulate algorithm use over time and reduce the better-than-average-effect. Participants will predict the job performance of 40 applicants based on their assessment information, either with or without algorithmic advice. Afterwards, participants will receive feedback on their *own* predictive accuracy, other people's accuracy, or no feedback. Two weeks later, participants will predict another set of 40 applicants and receive algorithmic advice. We hypothesize that participants who received accuracy feedback, especially on their own accuracy, will deviate less from the advice, make more accurate predictions, and become less overconfident. We ran a pilot study (N = 258) to investigate how different presentations of predictive validity estimates affect perceptions of hiring procedures, using a 2 (holistic vs. clinical synthesis) × 2 (visual and numeric display vs. numeric display only) × 2 (correlation vs. BESD) between-subjects design. The results revealed that participants are more likely to select an algorithm after they have seen that prescribed algorithms result in more accurate judgments than holistic approaches. Participants perceived larger accuracy differences between hiring procedures when accuracy was communicated numerically and visually, rather than only numerically.

Selection System Design: Exploring Predictor Choice and Predictor Importance for Validity and Diversity (13:25-13:50)

A. Susan M. Niessen[1], Fredrik Björklund[2], & Martin Bäckström[2]

[1]University of Groningen

[2]Lund University

What instruments are used in hiring procedures and how they are weighted affects validity and diversity. Many studies investigate how selection systems should optimally be designed to emphasize validity or diversity, but hardly any investigate how practitioners actually design selection systems for different goals. Therefore, in a pre-registered exploratory between-subjects study experiment, we investigated how *n* = 1212 people with hiring experience designed a hiring procedure, with the aim to: 1) maximize predictive validity, 2) improve the representation of female employees or 3) improve the representation of Black employees. Out of a list of nine common selection instruments, participants chose which instruments they wanted to include, and assigned a relative importance weight to each chosen instrument. First, compared to having a validity only goal, people who had a female representation goal included and weighted biodata somewhat more and cognitive ability and integrity tests somewhat less. People with a Black representation goal included and weighted biodata moderately more, and included and/or weighted cognitive ability tests, structured interviews, integrity tests, and emotional stability somewhat less. Some of these choices are in line with the empirical literature on adverse impact and predictive validity, but others are not. To estimate the effects of these different choices, the composite predictive validity and adverse impact for gender and ethnicity was estimated for each participant's selection procedure, based on a meta-analytic correlation matrix. Average composite validity and adverse impact for gender and ethnicity were virtually identical between conditions. So, even though professionals seem to make somewhat different choices when designing selection procedures for validity or diversity goals, that did not result in practically relevant differences in adverse impact or predictive validity.

The First Comprehensive Review of Resume Screening (13:50-14:15)

Isolde van der Schuur, Djurre Holtrop, & Janneke Oostrom
Tilburg University

Despite the enduring prevalence of resume screening in personnel selection, scientific research on its effectiveness remains limited. This literature review critically synthesizes existing studies on resume use, with particular attention to the validity of resume content, the impact of resume-based judgments, and the potential for bias. A central focus is placed on predictive validity: to what extent does resume screening predict job performance or other employment outcomes? An initial search identified 2,475 articles referring to resumes or curriculum vitae in their abstract; after data cleaning and exclusion of irrelevant uses, 534 records remained. These were screened by three independent reviewers, resulting in a final selection of 284 abstracts for analysis. Findings suggest a significant gap between the widespread use of resumes in practice and the lack of empirical support, particularly concerning predictive validity. Four records were identified that link resumes to job performance in non-academic contexts and show that resumes are a weak predictor of job performance.

The Implicit Trait Function (14:15-14:40)

Stijn Schelfhout, Femke Timmerman, & Eva Derous
Ghent University

An implicit trait policy (ITP) is a form of general domain knowledge (GDK), for which a worker's knowledge of effectiveness interacts with (trait-driven antecedent) dispositions like agreeableness. Through this interaction, an ITP succeeds to predict and explain effective work behavior. Research therefore uses an ITP as a validation mechanism in research that aims to predict and explain future work behavior like situational judgement tests (SJTs). Currently, ITP literature still debates the nature and the validity of the ITP concept, while also pointing towards methodological problems in the field. The present study introduces the implicit trait function (ITF) that transforms the dimension of a trait-driven antecedent disposition into the dimension of effectiveness, for both workers as well as behavioral responses. Proactive and retrospective applications of this function demonstrate that an ITP exhibits ample levels of construct-related validity, provided that the measuring instrument, the measurement scoring key, and the subsequent ITP calculations are properly facilitated. The ITF thus renders a comprehensive framework that not only integrates key contributions from the ITP literature but also addresses contemporary challenges within the field.

Identity in Reflection: Studying How Gender and Minority Status Shape Medical Students' Narrative Self-Evaluations in Comparison to Faculty Evaluations (15:00-15:25)

Marjolijn Wijnen, Laura Kalfsvel, & Karen M. Stegers-Jager
Radboudumc Health Academy, Nijmegen

When raters evaluate students' performances narratively, they systematically use different words for students from different demographic groups (i.e. ethnicity and/or gender). An alternative to these rater-based evaluations might be students' self-evaluations, as it would take the factor of the rater's prejudices out of the equation. People who are asked to reflect on their own performances, though, might also present themselves differently depending on which demographic group they belong to. In this research proposal, we present a methodology to determine whether self-evaluations would be a good alternative to rater-based evaluations for medical master students, specifically. In medical programmes, it is common practice to include self-evaluations in the performance portfolio that is used to make high-stakes decision on the student's progression. In our study, we will retrospectively investigate these self-evaluations to determine whether there are systematic differences between male and female, and ethnic majority and minority students. Using a large language model, we will compare the self-evaluations on valence (positive vs. negative), focus (communication vs. knowledge), and content (personal attributes vs. competencies). To further assess the suitability of self-evaluations for replacing rater-based assessments, we will also study whether gender and/or ethnicity-related differences in student's self-evaluations differ from those found in rater-based narrative evaluations.

Creating and Testing Stimulus Materials: A Comparison between Text- and Video-based Stimuli (15:25-15:50)

Aabidien Hassankhan, Laura Kalfsvel, & Karen M. Stegers-Jager
RadboudUMC Health Academy, RadboudUMC Nijmegen

Whenever research involves participants' exposure to stimulus materials, it is crucial that the latter are appropriate. If pre-existing stimuli are unavailable, designing and producing these materials oneself becomes crucial. However, this involves several complex compromises (e.g., realism vis-à-vis practicality). Furthermore, depending on the nature of the stimuli, this process can differ greatly. In medical education, stimulus materials are often necessary to controllably emulate educational processes. For a forthcoming study on the assessment of medical master students, we created written and video stimuli. We detail their creation and validation, and our decision-making process throughout. We also delve into the differences in producing written and video stimuli concerning the use of iteration, algorithmic testing, and expert opinion and show that researchers should be mindful of these differences. Our first results suggest that our efforts at generating written and video stimuli are successful. Written stimuli could be mass-produced and tested algorithmically, whereas video production necessitated the iteration of a handful of variants and more input from experts. We validated our stimuli through pilot testing, expert assessment and the comparison of subsequent descriptive data and effect sizes. However, how these stimuli perform in their target studies forms the ultimate litmus test.

An Interview with the Upcoming and Outgoing Organizers of the Dutch-Flemish Meeting on Selection Research (15:50-16:35)

Janneke Oostrom[1], Susan Niessen[2], Rob Meijer[2], & Marise Born[3]

[1]Tilburg University

[2]University of Groningen

[3]Erasmus University Rotterdam

The upcoming organizers of the Dutch-Flemish Meeting on Selection Research (Janneke Oostrom and Susan Niessen) will interview the outgoing organizers and creators of the Dutch-Flemish Meeting (Marise Born and Rob Meijer). They will reflect on the very successful past of our annual meeting and the bright future ahead of us!