Psychometrics at CoDE

Johannes Steinrücke[a]*, Stéphanie van den Berg[a], Remco Feskens[ab], Bernard P. Veldkamp[a]

* *Corresponding author. E-mail address: j.steinrucke@utwente.nl*

[a] Section of Cognition, Data and Education; Department of Learning, Data analytics and Technology; Faculty of Behavioural, Management and Social Sciences; University of Twente

[b] CitoLab, Cito, Arnhem, the Netherlands

Who is CoDE

The section CoDE (Cognition, Data and Education) is part of the LDT (Learning, Data analytics and Technology) department within the BMS (Behavioural, Management and Social Sciences) faculty at the University of Twente (UT).

CoDE has its roots in the educational sciences at the University Twente. In the early 1980s the department of Onderwijskundig meten en data-analyse (OMD; educational measurement and data analysis) was formed. Later, the same acronym OMD was used for a new name: Onderzoeksmethodologie, Meetmethoden en Data-analyse (Research methodology, measurement methods and data analysis). When a group of human factor researchers joined the OMD group in 2020, the new name of CoDE was assumed.

CoDE now consists of an international team of ~50 teachers and researchers with various backgrounds such as psychology, educational science, mathematics, and computer science.

Teaching & research at CoDE

At CoDE, we teach courses on methodology, data analysis, and psychology. A large part of our portfolio involves service teaching of methodology, statistics and data science courses for all but one of the BMS bachelor programmes (psychology, communication science, international business administration and management society & technology). We also teach in the master programme Educational Science and Technology, the Psychology master programme and various courses at other UT faculties.

We are heavily involved in international and national educational assessment, such as PISA and different Dutch national educational assessment (PEIL programme), where we collect the data, analyse the data and write the reports (e.g. Heitink et al., 2023; Meelissen et al., 2023). Much of our research has always focused on how to analyse such data properly, relating to group differences (e.g. nations, sexes, paper-and-pencil vs digital assessment) and how to account for these in such a way that fair comparisons can be made between individuals; this research often

was conducted under the heading of differential item functioning (e.g. Glas & Jehangir, 2013), data imputation (e.g. Glas & Geerlings, 2009), multilevel item-response theory (e.g. Fox & Glas, 2001), or Bayesian covariance structure modelling (e.g. Fox et al., 2020).

We have a strong track-record in item response theory modelling, using both frequentist (e.g. van der Linden & Pashley, 2000) and Bayesian (e.g. Fox & Glas, 2001; Schwabe, Boomsma, and Van den Berg, 2017; Fox, 2024) approaches for estimation. Data types other than test items have also received a lot of attention in our group: response times (e.g. Fox & Marianti, 2016) and other types of collateral data are now being used to more accurately measure individual differences and enabling more accurate feedback in formative testing.

Psychometrics at CoDE

CoDE is a globally recognized department with a long tradition in psychometrics with emeriti like Wim van der Linden, Cees Glas, and Theo Eggen. Traditionally, with its roots in educational measurement, psychometric research at CoDE has always had, and still has, a strong applied focus: the methods we develop solve real-world problems. This engineering mindset fits very well at a technical university like UT, with its *High Tech, Human Touch* philosophy.

Applied psychometrics

At CoDE we always played a crucial role in making psychometrical models actionable, by not only focussing on theoretical developments within the field, but also by applying our knowledge to use cases from various fields. For instance, applied psychometric work of CoDE includes item exposure control in the Law School Admission Test (LSAT; Van der Linden & Veldkamp (2007), the automated assessment of concept maps in the context teaching photosynthesis (Kroeze et al, 2021), using item response theory for variance decomposition to understand heritability of certain characteristics in quantitative genetics (van den Berg et al., 2007), using IRT to harmonise various data sets across different research centres (Van den Berg et al., 2014; Jovic et al., 2024) or combining item response and generalizability theory to better measure and compare teachers' instructional skills (Van der Scheer et al., 2017). In doing so, we ensure that our psychometric expertise and work results in real contexts, having direct impact in various domains.

Bayesian psychometrics

In the late 1990s, the Bayesian approach was adopted by Jean-Paul Fox, during his PhD project supervised by Cees Glas. Jean-Paul in turn supervised several new graduates. The multilevel Bayesian approach was also what drew Stéphanie van den Berg to Twente, who applied it in the context of behaviour genetics with her graduates (Van den Berg, Glas and Boomsma, 2007; Schwabe, Boomsma, and Van den Berg, 2017). In total, 7 theses on Bayesian psychometrics were finalised. Interestingly, the Bayesian psychometric modelling can be extended to

covariance structure modelling and multi-level modelling in general (dependency models; for instance Santos et al., 2021), so that our methods can be applied also outside the psychometric context, in fact any data set with a complex multilevel structure (for instance EEG and MRI data).

Role of ML and AI

Bernard Veldkamp was the first in Twente to introduce methods of data science into psychometrics. Text mining, large language models, and various machine learning methods are now commonly used in many new projects and PhD theses. The arrival of Maryam Amir Haeri also played a significant role in the generation of new ideas on how to apply machine learning methods effectively in psychometrics, and conversely, apply psychometric principles into machine learning. Currently, our PhDs are working on machine learning in test score equating problems, using reinforcement learning in (adaptive) testing, develop machine learning models for process data, and they are also exploring the potential of IRT in creating effective, transparent and explainable AI solutions. At CoDE we believe that the toolbox offered by the more traditional psychometric methods can be guiding for giving rise to these developments and can help ensuring that machine learning applications do not (further) become a black box procedure.

Who do we work with?

Most of our methods have applications in the domains of education and health, including mental health and psychopathology. Our most important partners are Cito, and several hospitals in the east of the Netherlands (MST, ZGT, Rijnstate, Isala and Deventer). More recently, psychometricians from CoDE started working with the iVTG (interuniversity progress test medicine). Selected other partners we work with are Carmel college, Auris, Oberon, Vrije Universiteit Amsterdam, Università di Bologna and more recently Universitetet i Stavanger. We further collaborate with various sections and departments across the University of Twente and receive guest researchers from countries all over the world, such as China, Italy, Turkey or Kazakhstan.

(Current) activities of CoDE

Projects

Researchers from the CoDE department are currently involved in a variety of (inter)national psychometric projects. For example, in the **AILIT** project, together with the Universitetet I Stavanger, a team from CoDE is developing recommendation systems in an international digital writing network aiming to improve writing motivation. Other researchers from CoDE are involved in analysing the data the 2023 edition of the **TIMMS** survey and were involved in the 2022 edition of **PISA** in the Netherlands. In the **HUMAN** project we work together with Universitetet I Stavanger

and Universitetet i Bergen to make reading assessments more enjoyable by modelling the attractiveness of questions and student preferences to further optimise assessment quality.

Conferences & symposia

Every year since 1984 we organise the **IRT workshop**, where psychometric researchers from mostly Dutch universities and institutes present and discuss their work on psychometrics in an informal setting. Aside from our own IRT workshop, our colleagues are regular attendees at, for instance, the **FREMO** biennial conference organised by the Centre for Educational Measurement (CEMO) at the University of Oslo, the **IACAT** annual meeting of the international association for computerized adaptive testing, the **NCME** annual meeting of the National Council on Measurement in Education in the USA, the **IMPS** annual meeting of the Psychometric Society or the **ITC** biennial meeting of the international test commission.

Non-research activities

Quite often, members of CODE are involved as psychometrics experts in projects both in the Netherlands and abroad. They serve in advisory boards of research projects in the Netherlands and Norway, are asked by the Dutch government for help, serve in technical advisory boards of testing agencies, contribute to lifelong learning programs of, for example, general practitioners, or work together with companies and/or institutes in contract research to solve problems these parties face in practice.

As quantitative enthusiasts at a social sciences faculty, naturally we play an important role in the statistical counselling of colleagues. We can therefore often be found as co-author on articles in non-psychometric journals on wide-ranging topics such as healthcare, education or psychology. Through our Methodology Shop we also offer methodological and statistical counselling to Bachelor and Master students, and PhD candidates of the BMS faculty.

CoDE's vision on psychometrics

At CoDE we are convinced that the future of psychometrics lies in the alignment of statistical approaches to methods from data science and machine learning to solve psychometric problems *and* machine learning problems. The central problem in psychometrics is the sound measurement of individual differences and the fair comparison of individuals, which is highly relevant also in the context of AI-based decision support systems. Measurement can be more reliable and efficient by incorporating not only classic items on a test but also involving many alternative types of data such as text, images and response times (e.g. Veldkamp, 2023). Integrating such different data types requires an efficient data pipeline and sound models. We are convinced that the field of psychometrics should go beyond the fantastic but limited realm of item-response theory models, such that we can solve problems where these models are no

longer valid. In turn we feel that item-response theory and psychometric reasoning have much to offer to the world of computer science, for instance in increasing explainability, fairness and transparency, of AI solutions.

References

Fox, J.P., & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika*, *66*(2), 271–288. https://doi.org/10.1007/bf02294839

Fox, J.P., Koops, J., Feskens, R., & Beinhauer, L. (2020). Bayesian covariance structure modelling for measurement invariance testing. *Behaviormetrika*, *47*(2), 385–410. P

Fox, J.P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*(4), 540–553. https://doi.org/10.1080/00273171.2016.1171128.

Fox, J.P. (2024). Redefining item response models for small samples. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/10769986241269886

Glas, C.A.W., & Geerlings, H. (2009). *A Study of Structural Modeling Using Plausible Value Imputation*. (LSAC Research Report Series; No. 08-03). Law School Admission Council.

Glas, C.A.W., & Jehangir, K. (2013). Modeling country-specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 97-115). (Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences). Chapman and Hall/CRC.

Heitink, M.C., Luyten, H., Meelissen, M.M., Veldkamp, B.P., van Langen, A. Keuning, J., Noordhof, R. (2023). *Digitale Geletterdheid in het basisonderwijs. Technisch rapport Peil.Digitale Geletterdheid 2022*. Universiteit Twente. https://www.onderwijsinspectie.nl/binaries/onderwijsinspectie/documenten/rapporten/2024/03/15/technisch-rapport-peil.digitale-geletterdheid-van-het-onderzoeksconsortium/

Jović, M., Haeri, M.A., Whitehouse, A., & Van den Berg, S.M. (2024). Harmonizing the CBCL and SDQ ADHD scores by using linear equating, kernel equating, item response theory and machine learning methods. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1345406

Kroeze, K.A., Van den Berg, S.M., Veldkamp, B.P., & De Jong, T. (2021). Automated assessment of and feedback on concept maps during inquiry learning. *IEEE Transactions on Learning Technologies, 14*(4), 460–473. https://doi.org/10.1109/tlt.2021.3103331

Meelissen, M.R.M., Maassen, N.A.M., Gubbels, J., van Langen, A.M.L., Valk, J., Dood, C., Derks, I., In 't Zandt, M., & Wolbers, M. (2023). *Resultaten PISA-2022 in vogelvlucht*. Universiteit Twente. https://doi.org/10.3990/1.9789036559461

Santos, J.R.S.D., Azevedo, C.L.N., & Fox, J.P. (2021). Bayesian longitudinal item response modeling with multivariate asymmetric serial dependencies. *Journal of Statistical Computation and Simulation, 92*(3), 488–523. https://doi.org/10.1080/00949655.2021.1965604

Schwabe, I. Boomsma, D.I., Van den Berg, S.M. (2017) Increased environmental sensitivity in high mathematics performance. Learning and Individual Differences, 54, 196-201.

Van den Berg, S.M., Glas, C.A.W., & Boomsma, D.I. (2007). Variance decomposition using an IRT measurement model. *Behavior Genetics, 37*(4), 604–616. https://doi.org/10.1007/s10519-007-9156-1

Van den Berg, S.M., De Moor, M.H.M., McGue, M., Pettersson, E., Terracciano, A., Verweij, K.J.H., Amin, N., Derringer, J., Esko, T., Van Grootheest, G., Hansell, N.K., Huffman, J., Konte, B., Lahti, J., Luciano, M., Matteson, L.K., Viktorin, A., Wouda, J., Agrawal, A., . . . Boomsma, D.I. (2014). Harmonization of Neuroticism and Extraversion phenotypes across inventories and cohorts in the Genetics of Personality Consortium: an application of Item Response Theory. *Behavior Genetics*, *44*(4), 295–313. https://doi.org/10.1007/s10519-014-9654-x

Van der Linden, W.J., & Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. In *Springer eBooks* (pp. 1–25). https://doi.org/10.1007/0-306-47531-6_1

Van der Linden, W.J., & Veldkamp, B.P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32*(4), 398–418. https://doi.org/10.3102/1076998606298044

Van der Scheer, E.A., Glas, C.A.W., & Visscher, A.J. (2017). Changes in teachers' instructional skills during an intensive data-based decision making intervention. *Teaching and Teacher Education, 65*, 171–182. https://doi.org/10.1016/j.tate.2017.02.018

Veldkamp, B.P. (2023). Trustworthy artificial intelligence in psychometrics. In *Methodology of educational measurement and assessment* (pp. 69–87). https://doi.org/10.1007/978-3-031-10370-4_4