2nd Scandinavian Applied Measurement Conference

June 8-10, 2026

Scandic Kødbyen, Copenhagen

https://biostat.ku.dk/SAMC/

Danish Data Science Academy

DSTS
DANISH STATISTICAL SOCIETY

Kristianstad University Sweden

Linnæus University

Swedish Defence University

Welcome message

The Swedish Network for Psychometrics and Metrology in the health sciences (PMhealth) was initiated in 2013 and has offered annual workshops. Following the 2022 PMhealth workshop, an organizing committee was formed to begin planning for an international conference: the Scandinavian Applied Measurement Conference (SAMC).

Similar to the PMhealth workshops, SAMC was intended to facilitate interaction and knowledge exchange among researchers, practitioners, and students. The SAMC theme is applied measurement, focusing on rating scales and other category-based measurement and assessment instruments.

Papers presented at SAMC 2024 concerned development, evaluation and quality assurance using Rasch measurement theory and related approaches, with an emphasis on the role of the individual person in the measurement process and analysis, as well as in the use and interpretation of results. The conference was multidisciplinary covering fields such as the health, social, educational, and behavioral sciences. SAMC 2024 had a strong international presence. Beyond presentations from Scandinavia, it hosted presentations from three continents, and a total of 10 countries represented.

Copenhagen now welcomes you to SAMC2026. This has been made possible by sponsorship from the Danish Statistical Society (DSTS).

Organizing Committee:

Prof. Peter Hagell, Kristianstad University, Kristianstad, Sweden (Chair)

Prof. Karl Bang Christensen, University of Copenhagen, Copenhagen, Denmark

Dr. Jeanette Melin, Swedish Defence University, Karlstad, Sweden

Prof. Kristofer Årestedt, Linnaeus University, Kalmar, Sweden

Assoc. Prof. Jørgen Holm Petersen, University of Copenhagen, Copenhagen, Denmark

Monday June 8	Tuesday June 9	Wednesday June 10
	08.30–10.00 Session 2	08.30–10.00 Session 5
10.00–12.00 Pre-conference workshop	10.00–10.30 Fika	10.00–10.30 Fika
Pre-conference workshop	10.30–12.00 Session 3	10.30–12.00 Session 6
12.00–13.00 Lunch	12.00–13.00 Lunch	12.00–13.00 Lunch
13.00–13.15 Opening	13.00–13.40 Poster	13.00–14.30 Session 7
13.15–14.30 Keynote presentation	Pitch	
	13.40-15.00 Posters	
14.30–15.00 Fika	and Fika	14.30–15.00 Fika
15.00–16.30 Session 1	15.00–16.30 Session 4	15.00–15.30 Closing
	Dinner **Conference dinner**	

SAMC Keynote presentation

Monday June 8

Rasch trees & Co: Combining Psychometrics and Machine Learning for the Detection of Differential Item Functioning

Carolin Strobl
Psychologisches Institut, Universität Zürich, CH

Rasch trees are a method for detecting differential item functioning (DIF) within the framework of model-based recursive partitioning. This framework enables the identification of differences in the parameters of psychometric models across two or more groups of persons. Its methodological roots lie in classification and regression trees developed in the field of machine learning. This presentation reviews the rationale of model-based recursive partitioning in general, with a particular focus on its application to Rasch and non-Rasch item response theory (IRT) models. It illustrates how Rasch trees and their extensions to non-Rasch IRT models can be used to detect both uniform and non-uniform DIF in two or more groups of persons. An important property of tree-based methods is that the groups do not need to be specified a priori but are learned from the data and can include interactions among multiple covariates. The presentation also addresses related issues that are relevant for interpreting the results of tree-based and other DIF detection methods, including the assessment of stability, the incorporation of effect sizes, equivalence testing,

SAMC Session 1

Session 1

Mariusz Grzeda
Fabio LaPorta
Serena Caselli

Session 1, 15.00-15.20

Measuring Quality of Life in Huntington's Disease: Key Challenges and Proposed Solutions

Mariusz Grzeda, Galen Research, UK

Objectives

Measuring quality of life (QoL) in Huntington's disease (HD) presents significant challenges. Progressive cognitive decline often limits patients' ability to self-report, making longitudinal assessment difficult and potentially incomplete. Proxy reporting, where caregivers or family members provide evaluations, offers a practical alternative, but introduces systematic bias. Evidence suggests proxies tend to rate patients' QoL more critically than patients themselves, which complicates interpretation and comparability. These issues raise fundamental questions about how to maintain conceptual equivalence and measurement accuracy when perspectives differ. This study addresses these challenges by exploring methodological strategies, grounded in Rasch Measurement Theory (RMT), to ensure proxy assessments remain aligned with patient perspectives and support meaningful measurement across disease stages.

Methods

Caregivers of individuals with Huntington's disease were recruited across Germany, the UK, and Italy to complete a proxy questionnaire developed in parallel with a patient-reported QoL instrument. Both versions were grounded in a shared conceptual framework derived from qualitative interviews and refined through translation, lay panel input, and cognitive debriefing. To address the methodological challenges, Rasch-based approaches were applied to optimize item functioning for proxy assessments while preserving conceptual equivalence, to account for systematic differences between patient and proxy judgments, and to explore strategies that position both perspectives on a common measurement continuum. These methods were designed to mitigate proxy-related bias and enable continuity in QoL measurement throughout disease progression.

Session 1, 15.00-15.20

Results
Analyses demonstrated that Rasch-based methods can successfully address key challenges in proxy measurement. Item-level refinements improved conceptual alignment, and linking strategies allowed proxy and patient perspectives to be placed on a shared continuum, supporting comparability across disease stages. These findings highlight the feasibility of maintaining measurement integrity despite systematic differences in judgments.
Conclusions
Measuring QoL in HD requires innovative solutions to overcome cognitive limitations and proxy-related discrepancies. Rasch-based approaches provide a robust framework for developing proxy instruments that support accurate, longitudinal, and patient-centered evaluation across all stages of the disease. By addressing these challenges, researchers and clinicians can ensure that QoL remains a central component of care and decision-making in HD.

Session 1, 15.20-15.40

Brain Injury Rehabilitation Trust Personality Questionnaires (BIRT-PQs) to measure neurobehavioral disability in patients with acquired brain injury: development of the short forms and cut-offs for their clinical interpretability

Fabio La Porta, IRCCS Istituto delle Scienze Neurologiche di Bologna, IT

Introduction: Brain Injury Rehabilitation Trust Personality Questionnaires (BIRT-PQs) are five questionnaires (150 total items) assessing neurobehavioral functions (motivation, emotional regulation, social cognition, disinhibition, and impulsivity) for patients with acquired brain injury (ABI). Given their length and resulting respondent's burden, developing short forms (SFs) was mandatory.

Objectives: 1) To develop SFs of the BIRT-PQ (SF-BIRT-PQ) using Confirmatory Factor Analysis (CFA) and Rasch Analysis (RA); 2) to make the scores clinically interpretable by calculating cutoff for each SF-BIRT-PQ to distinguish between behaviors that deviates (D+) or do not deviate (D-) from the mean of healthy individuals; 3) To make the SF-BIRT-PQ clinically usable by designing a digital ruler for calculating total scores and facilitate their clinical interpretation.

Session 1, 15.20-15.40

Methods: A multicenter, cross-sectional study included 154 ABI subjects and their caregivers (giving 308 records). CFA was used to study unidimensionality (UN) and local dependency (LD). RA fully assessed internal construct validity (ICV: monotonicity, UN, LD, invariance, DIF), reliability, and targeting was used to build SFs through item deleting. Cutoffs indicating D-/D+ patients were calculated using Z-scores developed from a RA conducted on 120 healthy subjects. Finally, a practical ruler (a nomogram) was developed to allow the clinical interpretation of the SF-BIRT-PQ measures.

Results: CFA showed lack of UN e LD. RA showed insufficient ICV and reliability. After rescoring 18 items and deleting 75 out of 150 items, the resulting five SF-BIRT-PQs showed adequate ICV with reliability consistent with single person measurement. Conversion tables were provided to obtain interval measures. The following D+ cutoffs were devised: SF-BMQ≥25 points, SF-BREQ≥19 points, SF-BSCQ≥14 points, SF-BDQ≥15 points, and SF-BIQ≥18 points. A ruler was developed for each questionnaire transforming raw item scores into measures that are interpreted based on cutoffs that identify D+.

Conclusions: After extensive modifications, including a 48% item reduction, five SF-BIRT-PQs were devised meeting the measurement requirements of the Rasch model were devised. The calculated cutoffs allowed interpretability of the SF-BIRT-PQ by quantifying the extent of the patient's behavioral deviation from that of an average healthy individual. Electronic rulers for each SF-BIRT-PQ provide several functions that greatly facilitate their administration and interpretation.

Session 1, 15.40-16.00

Advancing from Item-Level to Total Score-Based diagnosis in the Coma Recovery Scale–Revised: Evidence from a Diagnostic Accuracy Study

Serena Caselli
Azienda Ospedaliero Universitaria di Modena, IT

Session 1, 15.40-16.00

Introduction: The Coma Recovery Scale-Revised (CRS-R) is the gold standard for diagnosing persons with Disorders of Consciousness (PwDOC). Five out of six items provide scores linked to Unresponsive Wakefulness Syndrome (UWS), Minimally Conscious State (MCS), and emergence from MCS (eMCS), whereas no diagnostic criteria are associated with the total score (TS). Recently, a measurement-based study using Rasch analysis compared four sets of item-level diagnostic criteria [1].

Objective: 1) To define CRS-R TS cutoffs with the highest possible diagnostic accuracy (DA) based on the item-level criteria proposed by Caselli [1], applying both standard (a higher diagnostic criterion in one item is sufficient to assign the higher category) and alternative rules (at least 50% of items must support a given diagnostic criterion); 2) to compare the DA of TS cutoffs with previously published item-level cutoffs and to test their external validity.

Methods: A total of 380 PwDOC with acquired brain injury were included (giving 727 CRS-R assessments). DA of TS cutoffs derived from Caselli's item-level criteria [1], using both standard and alternative rules, was calculated. External validity was examined through discriminant validity analyses assessing differences in Disability Rating Scale (DRS), Glasgow Outcome Scale (GOS), and Levels of Cognitive Functioning (LCF) total scores across diagnostic categories (UWS, MCS, eMCS) using Kruskal–Wallis tests with post-hoc comparisons. Categories were defined using TS cutoffs. Agreement between TS- and item-based diagnoses from other authors was assessed using Krippendorff's alpha.

Results: All TS cutoffs derived using the alternative rule (MCS– = 9 points; MCS+ = 12 points; eMCS = 18 points) achieved the highest diagnostic accuracy (>93.5%) and showed a more balanced distribution across CRS-R distinct levels of performance ability. Significant differences in DRS, GOS, and LCF scores were observed across diagnostic categories for TS-based classifications. Krippendorff's alpha indicated high inter-author agreement among diagnostic approaches ($\alpha = 0.884$; 95% CI = 0.874–0.895).

Conclusion: While TS cutoffs based on the standard rule only partially addressed limitations of item-level criteria, the alternative rule enabled reliable total score-based diagnostic cutoffs with reduced measurement error, supported by strong external validity and inter-author reliability.

References:

1. Caselli S, et al. 10.1016/j.apmr.2024.12.009

SAMC Session 2

Session 2

Mark Elliott
Mike Horton
Marianne Müller

Session 2, 8.30-8.50

The effects of systemic and random rater behaviours on infit and outfit statistics under the MFRM and extended MFRMs

Mark Elliott
Department of Computer Science & Technology, University of Cambridge
Cambridge University Press & Assessment
UK

Infit and outfit mean square statistics (Wright & Masters, 1982) are residual-based fit statistics for Rasch models that may be used to evaluate the fit of any individual element of any facet of a model; within the many-facet Rasch model (MFRM; Linacre, 1994), this may include raters as well as persons, items and thresholds. There are two types of rater behaviour that may influence residuals. The first type of behaviour relates to on-trait differences between raters – different internalisations of the locations of thresholds while conceptualising the trait in the same way, which can manifest themselves in non-uniform rater behaviours such as central rating tendency or extreme rating tendency. The second type of behaviour relates to off-trait differences: the introduction of other dimensions to a rater's judgments and erratic rating, both of which may be considered sources of random noise within the context of a unidimensional measurement model.

We explore, via a simulation study, the effects of systemic rater behaviour – central and extreme marking tendency – and random behaviour on infit and outfit statistics for raters, items and thresholds, and how they interact when co-occurring, using first the standard, MFRM then extended MFRMs (Elliott & Buttery, 2022) that model non-uniform rater behaviour.

Session 2, 8.30-8.50

We find that under the MFRM, which does not separate systemic and random behaviours within residuals, infit and outfit statistics conflate the two types of behaviour, affecting their interpretability, and that this extends to fit statistics for items and thresholds as well as those for raters. Under appropriate extended MFRMs that isolate the relevant behaviours, however, systemic effects are accounted for; this results in more readily interpretable and useful infit and outfit statistics, which extends to fit statistics for items and thresholds as well as for raters.

References

Elliott, M. & Buttery, P. J. (2022) Extended rater representations in the many-facet Rasch model. Journal of Applied Measurement, 22 (1), 133–160.

Linacre, J. M. (1994). Many-Facet Rasch Measurement. MESA Press.

Wright, B. D., & Masters, G. N. (1982). Rating Scale Analysis. MESA Press.

Session 2, 8.50-9.10

Inverse dependency – identifying criterion values for lower bound Q3 residual correlations

Mike Horton, University of Leeds, UK

The assumption of local independence is central to all item response theory (IRT) models. The concept of local dependence (LD) is well-known, and previous studies have produced guidelines regarding the identification and interpretation of LD within an item set [1,2]. These studies investigated the critical values of the Q3 statistic commonly used to identify LD, and how the null distribution of these residual correlations is influenced by various factors. However, these investigations have exclusively focused on the highest observed Q3 values, which is where LD between items is identified.

inherent negative bias of the residual correlations and an unclear interpretation mechanism. However, through various real-world data examples, it has become apparent that the lowest observed Q3 values may indicate pairwise item incompatibility issues (i.e. items relating to things that do not typically occur at the same time, such as 'weight gain' and 'weight loss'). This 'inverse dependency' indicator may provide useful information to scale developers, but there are currently no guidelines as to any criterion lower bound Q3 values that might indicate an anomaly within an item set.
We therefore aim to replicate the previous large-scale Monte Carlo simulation study that investigated different factors that can influence the null distribution of residual correlations [2], but with the emphasis on identifying critical values for the lower bound Q3 values (signifying inverse dependency) rather than the higher bound Q3 values (signifying LD). The results of this study will be presented, with the objective of proposing guidelines that researchers and practitioners can follow when making decisions about inverse dependency during scale development and validation.

1. Andrich D, Humphry SM, Marais I. Quantifying local, response dependence between two polytomous items using the Rasch model. Applied Psychological Measurement. 2012 Jun;36(4):309-24.
2. Christensen KB, Makransky G, Horton M. Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. Applied Psychological Measurement. 2017 May;41(3):178-94.

Session 2, 9.10-9.30

Rasch Models for Incomplete Answers

Marianne Müller, Bern University of Applied Sciences, CH

Incomplete answers can occur for various reasons in measurement models, and the extent of nonresponse is greater in some research areas than in others. Most methods to handle missing data require that the missing data mechanism is MCAR (missing completely at random) or MAR (missing at random). Item nonresponses that are MCAR or MAR are called ignorable.
If the missingness depends on unobservable but important other variables in the study, the nonresponses are called nonignorable. For example, respondents may skip items due to lack of knowledge ("don't know"), unwillingness to answer (e.g., embarrassment), or some questions do not apply to them.
Multidimensional item response theory (MIRT) models directly estimate the tendency to omit items in addition to the usual person's ability. These models can deal with nonignorable nonresponses. However, due to their complexity they are not very widely used.

Methods

This paper examines a two-dimensional model for incomplete answers which belongs to the Rasch family of models. The model has been developped by Bertoli-Barsotti and Punzo (2013). They called it Rasch-Rasch model (RRM). One dimension provides information about the omitting behavior, the second dimension is related to the person's ability. Only dichotomous items are studied. Because the model is a member of the exponential family conditional maximum likelihood estimation and conditional likelihood ratio tests can be used. A simulation study was conducted to compare parameter estimates using the RRM and the Rasch model (RM).

Results

Session 2, 9.10-9.30

The simulations show slightly better values for bias, RMSE, and standard errors with RRM compared to RM using the package eRm in R and RUMM2030. RRM is much faster than eRm. A real data example illustrates the potential of the RRM.

Discussion

The RRM is suitable for data with nonignorable nonresponses. If there are a lot of missing values, it could be an alternative to eRm, which becomes very slow in these situations. The model also allows to estimate the tendency of an item to be skipped by respondents.

References

Bertolli-Barsotti,L. & Punzo,A. (2013). Modelling missingness with a Rasch-type model. Publications de l'Institut de Statistique de l'Université de Paris. 54(1-2): 29-44.

SAMC Session 3

Session 3

Jacob B. Jørgensen
Ann-Sophie Buchardt
Leonardo Pellicciari
Carolina Fellinghauer

Session 3, 10.30-10.50

Assessing internal construct validity of SDAI and CDAI in rheumatoid arthritis: A study using confirmatory factor analysis, minimal clinically important difference and other psychometric tests

Jacob B. Jørgensen
Copenhagen Center for Arthritis Research
Center for Rheumatology and Spine Diseases, Rigshospitalet, DK

Session 3, 10.30-10.50

Background

Rheumatoid arthritis (RA) disease activity is commonly monitored using the Simplified Disease Activity Index (SDAI) and the Clinical Disease Activity Index (CDAI). Both indices include tender and swollen joint counts and patient and physician global assessments, while SDAI additionally incorporates inflammation. Psychometric studies have primarily addressed correlations with other measures, whereas evidence of latent traits remains limited. Structural equation modeling (SEM) allows evaluation of whether these indices adequately represent the underlying construct of disease activity. Additional key properties include internal consistency reliability and minimal clinically important difference (MCID).

Objectives

To evaluate internal construct validity, internal consistency reliability, and MCID of SDAI and CDAI using longitudinal routine care data from patients with RA.

Methods

This study included adult patients with RA from the Danish DANBIO registry initiating their first biologic or targeted synthetic disease-modifying antirheumatic drug between 1999 and 2019 and had SDAI and CDAI recorded at baseline and 6 months. Item relationships were examined using correlations, pairwise scatter plots, and histograms to visualize overall patterns, and dimensionality was assessed with scree plots. Models were fitted for baseline and follow-up separately and pooled for SDAI and CDAI. Model fit was evaluated using the comparative fit index (CFI >0.95), supported by χ^2 and RMSEA. Sensitivity analyses included stratification by age and sex, exclusion of outliers and correlated tender/swollen joint errors. Anchor-based MCIDs were estimated using ROC curves and Youden's index, with patient-reported change on a 7-point Likert scale as anchor. Internal consistency reliability was assessed using McDonald's ω.

Results

Among 5,200 patients, scree plots supported one-dimensionality for both indices. Cross-sectional models demonstrated comparable item loadings and acceptable fit whereas

Session 3, 10.30-10.50

longitudinal models showed poor fit. MCIDs for improvement were −5.48 for SDAI and −7.50 for CDAI, and for worsening −2.30 and −5.50. McDonald's ω indicated good reliability (0.80–0.84).

Conclusions
SDAI and CDAI demonstrated adequate internal construct validity and internal consistency reliability in a large RA cohort, supporting their clinical use. However, longitudinal construct validity was limited. Identified MCIDs required greater change to detect improvement than worsening. Further studies should compare more indices and assess cross-national validity.

Session 3, 10.50-11.10

Severe chronic disease and school well-being- a Danish nation-wide cohort study

Ann-Sophie Buchardt
Mary Elizabeth's Hospital, Rigshospitalet, DK

developed by expert group commissioned by the Ministry of Education to monitor well-being among all Danish public school students on a yearly basis. A prerequisite for a survey to be considered an adequate measurement instrument is a unidimensional structure as demonstrated by a modern test theory (MTT) model, meaning that each (sub)scale reflects one construct (e.g., loneliness). However, the structural validity of DSWQ has only been sparsely evaluated with MTT on a 2015 cohort of school children. This study investigated the dimensionality and measurement properties of the survey contributing to clarity on interpretation and application with a focus on students with severe chronic disease (SCD). Data were extracted from various Danish national registers, including health registers and data from Statistics Denmark. The study population included live-born children born in Denmark starting in school between 2005 and 2023 (without prior migration). Students with SCD were identified using ICD-10 codes. Four subsamples – female and male, healthy and chronically ill – were assessed by confirmatory factor analysis (CFA) and Rasch analysis. CFA

Session 3, 10.50-11.10

model fit was evaluated using the chi-squared statistic and indices of close fit. Rasch fit was evaluated with item fit statistics. Both a one-factor solution and two-factor solutions with scores based on two separate subscales were considered for each subsample.

Session 3, 11.10-11.30

Development of short forms for the Unified Balance Scale (UBS): a Rasch analysis study

Leonardo Pellicciari
IRCCS Istituto delle Scienze Neurologiche di Bologna, IT

Introduction: The Unified Balance Scale (UBS) is a 27-item scale developed using Rasch analysis (RA) by merging three instruments assessing balance (Berg Balance Scale, Tinetti Scale, Fullerton Advanced Balance Scale) within the neurological rehabilitation setting [1].

Objective: To develop two UBS short forms (UBS-SF) of increasing difficulty to reduce administrative and patient's burden.

Methods: An UBS short form (16 items) was calibrated using the original UBS dataset [1]. This scale was administered to 502 patients with balance disorders of different neurological etiologies. A validation sample (VS) and a confirmatory sample (CS), each consisting of 251 subjects, were randomly created. RA was used to examine internal construct validity (ICV: invariance, monotonicity, unidimensionality, local independence, DIF), reliability, and targeting in each sample. After achieving adequate ICV, two SF (UBS-SF1 and UBS-SF2) of increasing difficulty were co-calibrated with the full scale and an algorithm was developed to choose the more targeted SF to the patient's ability.

Results: Baseline RA for VS revealed model misfit ($\chi2=64.248$, $p<0.0001$), and widespread violations of ICV requirements; reliability was adequate for single-subject measurements (PSI=0.946). After rescoring disorder items and creating 4 subtests accounting for local dependence, the data showed adequate model fit ($\chi2=32.327$, $p=0.22$), local independence, acceptable unidimensionality, negligible DIF, and adequate reliability for single-subject measurements (PSI=0.938). This final solution was replicated and anchored on the CS,

Session 3, 11.10-11.30

showing a stable calibration satisfying ICV and with reliability adequate for individual measurements (PSI=0.940).
Both an easier (UBS-SF1: 9 items; total score: 0-22 points) and a more difficult SF (UBS-SF2: 7 items, total score: 0-17 points) were co-calibrated with UBS-SF. A 14 points score on UBS-SF indicated a reduction in information for UBS-SF1 and an increase in information for UBS-SF2. Furthermore, the administration of three probing items (TB06, TB07, and BBS09) predicted with an accuracy>80% the most appropriate SF to be administered. A score >=14 on UBS-SF1 indicated the need to switch to UBS-SF2.

Conclusion: Two UBS short forms satisfying the Rasch model requirements were devised. A simple algorithm can guide clinicians to choose the more targeted short form to the patient's ability.

Session 3, 11.30-11.50

Measurement of the Experienced Burden with Secondary Condition in Spinal Cord Injury: Validation and True Score Equating for International Survey Integration

Carolina Fellinghauer, Swiss Paraplegic Research, CH

Background: Persons with Spinal Cord Injury (SCI) are at high risk of experiencing several secondary health conditions. These conditions have an important negative impact on a range of key health outcomes. At clinical level, the identification of secondary health conditions is part of routine data collection.
Objective: The purpose of the study is to investigate the psychometric properties of the SCI-SCS with data from N 22'732 participants from N = 35 countries that participated in one or both waves of the International Spinal Cord Injury survey (InSCI; 2017 and 2023).
Methods: The measurement properties of the SCI-SCS were investigated with the Partial Credit Model. True score equating and anchored analyses were conducted to assess

Session 3, 11.30-11.50

changes in item-parameter fit across countries, evaluate the suitability and limitations of establishing a common SCI-SCS metric for international use, and derive a sound and equitable approach for measuring the experienced burden of secondary conditions in a global assessment context.
Results: With only a few exceptions, countries showed good alignment with the common SCI-SCS metric. Challenging items were examined with respect to differences in perceived meaning and potential cultural factors that may influence response patterns.
Conclusion: The Rasch-based equating methodology enables the development of a single, internationally comparable summary measure of the burden of secondary conditions among individuals with SCI. This supports its application in epidemiological research and public health monitoring across diverse world regions.

SAMC Poster session

Poster session

Dolma
Dornonville de la Cour
Gjertsson
Modig
Hagell
Christensen
Morgan
Elliott
BushkePauli Kristensen

Poster session, 13.00-13.04

Servant Leadership as Degree or Configuration? A Rasch Measurement Approach

Serkan Dolma
Pamukkale University, TR

Servant leadership, originally articulated by Greenleaf (1977), refers to a leadership approach in which the leader's primary motivation is to serve others, emphasizing follower growth and well-being as well as responsibility toward the broader community beyond the organization. Among the various attempts to operationalize servant leadership, one of the most widely used instruments is the scale developed by Liden et al. (2008), which conceptualizes servant leadership in seven dimensions: emotional healing, creating value for the community, conceptual skills, empowering, helping subordinates grow and succeed, putting subordinates first, and behaving ethically.

Prior work has treated servant leadership both as a multidimensional construct and as a global, unidimensional construct, raising the question of whether these dimensions reflect variation in degree along a single servant leadership continuum or represent qualitatively different virtues that jointly define the construct. Within this literature, Liden et al.'s work reflects an inclination toward a degree-based interpretation, as evidenced by higher-order factor models and the development of a unidimensional short form (2014, 2015). However, the substantive definitions of the seven dimensions suggest that these leadership characteristics may not be straightforwardly ordered along a single continuum, but may instead reflect qualitatively distinct aspects of what it means to enact servant leadership. Against this background, the present study examines whether servant leadership can be meaningfully treated as a unidimensional construct within a unidimensional polytomous Rasch measurement framework.

Data will be collected from employees across multiple sectors, randomly assigned to one of three servant leadership instruments: (a) the original 28-item scale with four items per dimension, (b) a 7-item short form consisting of one item from each dimension, and (c) a vignette-based measure in which each dimension is represented by a brief description of a fictitious manager, with respondents making resemblance-based ratings of how closely the description matches their own manager. For the 28-item instrument, each dimension will first be analyzed as a separate subscale using a Rasch model; contingent on adequate fit, items within each dimension will be combined into higher-order polytomous items and reanalyzed. The short form and the vignette-based instrument will be examined directly using unidimensional Rasch models. By integrating evidence across the three instrument

Poster session, 13.00-13.04

formats, the study aims to clarify whether servant leadership is more appropriately interpreted as a global unidimensional construct or as a configuration of distinct leadership characteristics from a Rasch measurement perspective.

Poster session, 13.04-13.08

Rasch Analysis of the 7-Item Fatigue Severity Scale (FSS–7) for Adults 6 and 12 Months after Traumatic Brain Injury

Frederik Dornonville de la Cour
Copenhagen University Hospital - Rigshospitalet, DK

Introduction: Fatigue is a common and debilitating consequence of traumatic brain injury (TBI), yet validated self-report measures for this population are limited. The 7-item Fatigue Severity Scale (FSS–7) is recommended for stroke research, but has not been evaluated in TBI using modern test theory approaches. This study aimed to evaluate validity, reliability, and targeting of the Norwegian FSS–7 for adults with moderate-to-severe TBI.
Methods: Rasch analysis was conducted on longitudinal data from 94 adults with intracranial injuries, assessed at 6 and 12 months post-injury (N = 185). The mean age was 45.8 years (SD = 13.6), and 20% were female. We included the persons twice, as there was no evidence of longitudinal local dependence between responses at the two timepoints nor of longitudinal (DIF) for items at the second timepoint relative to the latent variable at the first timepoint. The Partial Credit Model (PCM) and graphical log-linear Rasch models (GLLRM) were used to test item fit, no local dependence (LD), no DIF and invariance across sex, age, education, and sleep disturbances (Epworth Sleepiness Scale and Insomnia Severity Index), as well as no DIF over time (i.e., no item parameter drift). Reliability was estimated using Hamon and Mesbah's Monte Carlo procedure, and targeting was assessed using test information indices.
Results: The FSS–7 did not fit the PCM due to misfit of item 4 and moderate LD between items 5 and 6. The FSS–6, excluding item 4, fit to a GLLRM accounting for the LD. No DIF or item parameter drift was detected. Reliability of the FSS–6 was excellent (r = 0.936). Targeting was suboptimal, with an average of 57% of the maximum obtainable test information reached.

Poster session, 13.04-13.08

Discussion: The FSS–7 is not valid for Norwegian adults 6 and 12 months post-TBI. The FSS–6 provides an essentially valid and reliable measure of fatigue for this population, supporting its use in research. The FSS–6 is invariant across sex, age, education, sleep disturbances, and time points (6 and 12 months post-injury). Suboptimal targeting may limit the clinical utility of the FSS–6, and future research should re-assess this.

Poster session, 13.12-13.16

Psychometric Evaluation of the Swedish Early Development Instrument: Assessing Developmental Health in Young Children

Sofia Gjertsson
Department of Public Health and Caring Sciences, Uppsala University, SE

Introduction
Early childhood is a period of profound advances, to a large extent shaped by the child's environment. High-quality early education significantly impacts and improves children's developmental outcomes. Identifying needs and vulnerabilities in early education becomes crucial for equitable possibilities for children to reach their potential. The Early Development Instrument (EDI) is a teacher-reported instrument, developed in Canada in the 1990s and used as a national census in both Canada and Australia for over 20 years. The EDI identifies vulnerabilities by assessing developmental health across five domains; a) physical health, b) social competence, c) emotional maturity, d) language and cognitive development, e) communication skills and general knowledge.
Aim
Following initial piloting in Sweden, this study aims to assess the psychometric properties of the Swedish translation of the EDI, using Rasch Measurement Theory.
Methods
Data will consist of EDI observations from at least 2000 children in four different municipalities, collected in April 2026. Each domain of the instrument will be analysed separately to assess unidimensionality using principal component analysis of residuals. Any local dependency will be analysed using Yen's Q3. Additionally, differential item functioning by gender will be examined.
Model fit will be examined using item and person fit measures, and item fit residuals.

Poster session, 13.12-13.16

Category functioning, including threshold ordering and category probability curves, will be assessed to evaluate response categories and participants' understanding of items. Finally, reliability will be estimated using person separation index.

Results and Discussion

Results will be presented on domain level. The analyses will indicate how the Swedish version of the EDI functions, specifically whether translation or cultural context has affected item performance or validity. The analyses are expected to support unidimensionality while highlighting item issues related to translation or item redundancy. The results will answer the question of validity of the instrument but also guide decisions regarding future studies with the instrument. For instance, any pitfalls related to scale up of the instrument will guide suitability for population-wide implementation – a potential step towards improving equitable health and development in young children.

Poster session, 13.16-13.20

Translation and cultural adaptation of the Self-Care of Hypertension Inventory (SC-HI): a cognitive interview study

Maria Modig
Unit for palliative medicine, Geriatric clinic, Kalmar County Hospital, Linnaeus University, Faculty of Health and Life Sciences
SE

Introduction: Self-care is crucial for individuals with hypertension, making it essential to assess accurately. This study aimed to translate and culturally adapt the Self-Care of Hypertension Inventory (SC-HI) into Swedish and to evaluate its content validity and response processes.

Methods: The study followed a two-phase process: (i) translation of the original English version in accordance with ISPOR principles and PRO Consortium guidelines, using forward and backward translation by four independent professional translators; and (ii) cognitive interviews with adults diagnosed with hypertension, including one participant with Swedish as a second language. Fifteen participants were recruited through convenience sampling from primary care and other community settings in southern Sweden. Data were analyzed

Poster session, 13.16-13.20

deductively using the Survey Response Model themes: comprehension, recallability, sensitivity/decision-making, and question response; through an iterative, reparative approach aligned with established cognitive interviewing practices.
Results: In the translation phase, differences in wording and phrasing were identified and discussed among the authors and the instrument's developer to ensure conceptual equivalence before changes were made. Participants generally perceived the SC-HI as easy to understand, with relevant items, clear instructions, and response categories well aligned with the questions. Interviews also informed judgments of content validity (i.e., content relevance and clarity) and revealed linguistic challenges requiring adjustments for better cultural fit. Terms such as "monitor" and "regularly" were replaced with more appropriate Swedish equivalents. Items addressing recognition of high blood pressure posed interpretation difficulties due to ambiguities in phrasing and response options. Additionally, some respondents considered one item about weight as sensitive. Revisions included splitting one item, clarifying instructions, and using fully labelled response options to improve clarity and usability.
Discussion: This study provide evidence for content validity and response processes, and offering transferable insights on item interpretation that may support future refinements of the original SC. Further psychometric evaluation is required before clinical and research use.

Poster session, 13.20-13.24

Rating scale quality assurance in the Swedish national quality registers

Peter Hagell, Kristianstad University, SE

Introduction: There are more the 100 Swedish national quality registries (NQRs) that systematically collect information on healthcare and treatment across various disease areas. The primary purpose of the NQRs is to monitor and improve care, but they also provide a valuable resource for research. An important component of these registries is the use of rating scales. Rating scales are used to assess various aspects of health and disease and play a central role in documenting and monitoring patients' health status and treatment outcomes, and are also used in clinical decision making. However, a systematic quality control of rating scales used in NQRs is currently lacking, which makes data quality uncertain and it is unclear how rating scale data can be reliably used in clinical practice and research. This project aims to document the quality of rating scales used in NQRs, starting

Poster session, 13.20-13.24

with the Parkinson's disease subregistry (ParkReg) of the Swedish Neuro Registries.

Methods: The work will be based on available registry data, which for ParkReg means 16 rating scales with about 400-12000 observations each. Analyses will comprise descriptive item level analyses, confirmatory factor analysis (to address the latent structures of the scales), Mokken scale analysis (to address ordinal properties), and Rasch measurement theory (to address linearity through, e.g., local independence, measurement invariance, and measurement uncertainty). Qualitative evaluations (e.g., relevance assessment, cognitive interviews) will be carried out as needed.

Results: Not yet available.

Discussion: Using national registry data allows for thoroughly evaluation of rating scales regarding their strengths and limitations in a representative real-world context. This will provide new insights into how rating scale data from NQRs can be reliably used in clinical practice and research. In addition, it will serve as guidance to propose evidence-based improvements where necessary. Quality-assured rating scales are essential for obtaining data that support reliable comparisons and enable detailed clinical monitoring, as well as providing high quality data for clinical studies. As such, this project has the potential to contribute to improved decision-making support, better care, more robust research findings, and a stronger evidence base for treatment recommendations.

Poster session, 13.24-13.28

Critical values for Rasch item fit statistics

Karl Bang Christensen, University of Copenhagen, DK

Item fit statistics are very frequently reported in Rasch analysis. Many of them have an unknown asymptotic distribution and their interpretation rely on rules-of-thumb. Extensive simulation studies have shown that it is not possible to provide a single critical value that will apply in all settings. This is a problem for all Rasch validation studies that relies on proprietary software. We introduce an easy-to-use software application that will help in establiblishing critical values on a case by case basis.

Poster session, 13.28-13.32

Team Psychological Safety Scale in the health care setting
– feasibility and assessment of psychometric properties using modern test theory in a Swedish translation
Proposal for a planned study

Sara Morgan
Institution for public health and caring sciences, Uppsala University, SE

Introduction
Psychological safety (PS)—the climate in which team members deem it safe to take interpersonal risks—is a key determinant of effective teamwork. In health care, PS is essential for patient safety, as it enables staff to safely surface mistakes and risks. As interest in PS has grown within health care research and practice, reliable and valid measurement of the construct has become increasingly important.
The Team Psychological Safety Scale (TPSS), developed by Amy Edmondson in the 1990s within a US corporate context, is widely regarded as the gold standard for assessing PS. The scale has since been adopted across a range of settings and translated into multiple languages.
Despite the widespread use, the psychometric properties of the TPSS have only been sparsely evaluated, and evidence regarding its functioning in health care settings and in translated versions remains limited. In particular, no studies using modern test theory were identified in our search. To support the valid use of a Swedish translation of the TPSS in health care contexts, a systematic psychometric evaluation is warranted.
Method
A digital survey with a Swedish translation of the 7 item TPSS and background factors including gender, age, profession and workplace will be distributed to nurses in child health centres (N≈250) in three regions in Sweden. A further sample will be collected recruiting a mix of primary health care staff through social media (N≈100).
Psychometric properties of the TPSS will be examined using Rasch Model Theory (RMT). Analyses will assess item functioning, dimensionality, and overall model fit to evaluate the appropriateness of summing item scores into a total scale score. Person-item distribution will be assessed to reveal any floor or ceiling effects, and internal consistency will be assessed using the Person Separation Index.

Poster session, 13.28-13.32

Discussion
As the scale was developed in a US corporate context, evaluating its psychometric performance on Swedish healthcare settings is important addition to the PS literature. RMT analyses may provide detailed insights into item functioning and scale structure, thereby providing novel material and a new piece to the puzzle of understanding the construct itself.

Poster session, 13.32-13.36

Parameterisation versus parsimony: The application of information criteria to the validation of model specification for polytomous and many-facet Rasch models

Mark Elliott
Department of Computer Science & Technology, University of Cambridge
Cambridge University Press & Assessment
UK

The choice of model specification for polytomous Rasch models between the RSM and the PCM is in the first instance a theoretical one, based on considerations of whether items should be expected to share the same threshold structure. However, empirical tests of whether the data support the chosen model specification form part of the validation of an instrument; if data do not support the chosen model specification, this may lead to decisions around whether to respecify the model or revise the instrument.

There are a range of metrics for selecting between competing model specifications, including Fisher's likelihood ratio test and information criteria based on different theoretical considerations such as AIC, and BIC, which explicitly penalise more highly parameterised models to different extents (Dziak, Coffman, Lanza, Li, & Jermiin, 2020), and which have differing biases towards more highly parameterised or more parsimonious models under different conditions.

In this study, we use simulated data to investigate the efficacy of different metrics for model selection. Firstly, we generate data sets according to the rating scale model (RSM) and the partial credit model (PCM) to determine the effectiveness of each metric under a range of conditions with different numbers of items, maximum scores and sample sizes. We apply the different metrics for model selection to determine the proportions of correct model selections based on the generating parameters under different conditions. We then repeat

the process for the many-facet Rasch model (MFRM) to select between competing extended rater representations: the global model and four different extended rater representations (Elliott & Buttery, 2022), which present questions of rater specification similar to those presented by the choice between the RSM and PCM.

From the analyses of the performance of the different metrics, we infer practical guidelines for empirical model testing.

References

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. Briefings in bioinformatics, 21(2), 553-565.
Elliott, M., & Buttery, P. J. (2022). Extended rater representations in the many-facet Rasch model. Journal of Applied Measurement, 22(1), 133–160.

Partners' experiences of low-risk induction of labor- From qualitative knowledge to quantitative scales

Camilla Bushke Pauli
Department of Obstetrics and Gynecology
Sahlgrenska University Hospital
Department of Obstetrics and Gynecology, Institute of Clinical Sciences
Sahlgrenska Academy, University of Gothenburg,
Gothenburg, SE

Background:
Induction of IOL (IOL) is a common obstetric intervention that almost 30% of all women and their partners will experience. However, research on experiential outcomes has primarily focused on the birthing person. Partners are key participants in the childbirth process, and their experiences may influence both the partner’s immediate and long-term wellbeing as well as medical outcome for mother and child. Despite this, there is a lack of theoretically grounded and psychometrically sound scales for measuring partners’ experiences of IOL.

Poster session, 13.36-13.40

Aim:
The overall aim of this doctoral project is to explore partners' experiences of IOL and to develop and psychometrically evaluate partner-reported experience of IOL scales using Rasch Measurement Theory (RMT).
Design and Methods:
The project comprises four interconnected studies, all conducted within the Swedish multicentre Outpatient Induction Trial (OPTION, EUCT-nr:2023-507164-39-00). The first two studies used qualitative methods to explore partners' experiences of IOL through in-depth interviews and free-text answers from the OPTION-questionnaires. Taken together, the qualitative studies found that partners experienceuncertainty and loss of control, struggle to manage responsibility and preparedness, as well as that contextual and relational conditions shape their experience.
The third study will develop scales based on construct theories to capture the IOL experiences identified in the first two studies. Specifically, the qualitative findings will inform item development and the formulation of a priori item hierarchies. At this stage, proposed partner-reported IOL experience scales will also be tested through cognitive debriefing. Subsequently, the fourth study will evaluate the psychometric properties of the IOL scales using RMT.
Significance:
By grounding the construct theory in the latest qualitative knowledge, we will develop items for the IOL experience scales that reflect what it means to go "from less to more" indifferent domains of IOL experiences among partners. Thus, the resulting scales should enable valid, reliable, interval-level measurement of partners' IOL experiences. This doctoral project addresses a methodological and clinical gap and contributes to advancing person- and family-centred measurement in maternity care.

Poster session, 13.40-13.44

Setting norms for psychometric instruments using item response theory

Bjarke Hautop Kristensen
University of Copenhagen, DK

When psychometric instruments are used for screening purposes, several components are important in the identification of useful normative values: (i) the availability of large representative data sets; (ii) content validity; (iii) psychometric validity of the instrument and its scoring. For instruments scored using raw sum scores, a typical approach is to compute empirical percentiles in a large representative data set. We propose that the use of item response theory models may improve the accuracy and consistency of the assessments between different populations. We do this using Bayesian IRT Burkner (2021).

References

Bürkner, Paul-Christian (2021). "Bayesian Item Response Modeling in R with brms and Stan". Journal of Statistical Software 100 (5). doi: 10.18637/jss.v100.i05

SAMC Session 4

Session 4

Fredrik Gasser
Karl Bang Christensen
Sara Scholtens

Session 4, 15.00-15.20

Person-centred measurement

Fredrik Gasser, Kristianstad University, SE

Introduction: Person-centred measurement (PCM) and closely related terms such as patient-centred measurement and person-centred metrology are increasingly encountered, explicitly or implicitly, in relation to category-based measurement in the health sciences. These terms (collectively referred to as PCM here) imply that the person is at the centre of measurement. However, it is unclear what is meant by putting the person at the centre. Here we review the conceptualisation and operationalisation of PCM and discuss their implications.
Methods: The paper is based on a review of relevant literature.
Results: Five main conceptualisations and operationalisations of PCM were identified: (i) measurement of patient perspectives on their health and care using patient-reported outcome measures (PROMs) and patient-reported experience measures (PREMs); (ii) involvement and consideration of the perspectives of target population representatives in the development of, e.g., PROMs; (iii) measurement of outcomes considered central to person-centred care, such as perceived wellbeing, satisfaction and involvement with care; (iv) the role of the person in the measurement process, where the person acts as a measurement instrument rather than the object of measurement; and (v) the meaning of measurement results and their applicability to individual persons.
Discussion: While there are overlaps between the identified views on PCM they also represent different viewpoints. These are critically discussed in terms of their implications from epistemological, clinical, practical and measurement perspectives. It is argued that there is a need for better agreement on the meaning of PCM, preferably across disciplines.

Session 4, 15.20-15.40

Visualizing measurement invariance using stratified CICC plots

Karl Bang Christensen, University of Copenhagen, DK

Communicating Rasch model fit is essential for establishing the empirical validity of Patient-Reported Outcome Measures (PROMs). While statistical tests for Differential Item Functioning (DIF) are common, graphical evaluations provide critical context regarding the magnitude and location of misfit. This study utilizes the R function CICCplot to generate Conditional Item Characteristic Curves (CICCs), which describe the expected item mean as a function of the total score. By stratifying these plots into subgroups researchers can visually inspect for DIF. The stratification of empirical curves allows for the comparison of subgroup performance directly against the model-based CICC. Using data from a Danish BREAST-Q validation study we illustrate how the impact of statistically detected DIF can be evaluated. Stratified CICCs extend beyond p-values to offer a qualitative understanding of item performance across populations yielding a visual approach that facilitates interpretation.

Session 4, 15.40-16.00

Improving Measurement Validity in Sexual Health Surveys: Insights from Cognitive Interviews on Gonorrhea and Chlamydia Risk Factors

Sara Scholtens, Public Health Agency, SE

This study presents findings from cognitive interviews conducted to evaluate and refine a survey instrument designed to assess behaviors and risk factors associated with gonorrhea and chlamydia among young adults in Sweden. The survey, part of a collaboration between the Public Health Agency of Sweden and regional infection control units, targets individuals aged 18 and above who have recently been tested for these infections.

Seven in-depth cognitive interviews were conducted with participants aged 17–32 (mean 24), including both genders and diverse educational backgrounds. The interviews revealed significant issues with question clarity, response options, and survey structure, particularly regarding the use of negations, ambiguous time frames, and culturally insensitive language. Key problems included misinterpretation of questions about condom use, sexual partners, and substance use during sex, as well as confusion about the relevance of certain questions for women and non-heterosexual individuals. Participants also highlighted the need for clearer definitions, more inclusive language, and logical question sequencing to reduce cognitive burden and measurement error.

Recommendations for revision include rephrasing questions to avoid negations, simplifying response scales, providing clear definitions for technical terms and restructuring the survey to improve flow and relevance. The revised instrument aims to enhance measurement validity and reduce bias, ensuring more accurate and reliable data collection for public health interventions.

The study underscores the importance of cognitive interviewing in identifying and addressing measurement issues in sensitive health surveys, particularly in multicultural and diverse populations. These insights are relevant for researchers and practitioners developing and validating measurement instruments in sexual health and related fields.

SAMC Session 5

Session 5

Samantha Ehrlich
Purya Baghaei
Farshad Effatpanah
Bengü Börkan

Session 5, 8.30-8.50

Examining Construct Validity of Large-Scale Mathematics Assessments Using Rasch and Explanatory Models

Samantha Ehrlich

Large-scale mathematics assessments are widely used to make inferences about students' mathematical knowledge; therefore, the validity of these inferences must be well supported by evidence. In this presentation, I demonstrate how construct validity can be examined from multiple perspectives using large-scale assessment data to evaluate how effectively mathematics items function as indicators of their intended construct. Drawing on Rasch measurement theory, I first assess whether a multi-format national mathematics test demonstrates sufficient coherence to justify treating total scores as measures of a single underlying attribute. I then use explanatory models to investigate the alignment between the intended and enacted constructs by analysing how item features related to cognitive demand and mathematical content are associated with empirical item difficulty. Finally, I explore a potential source of construct-irrelevant variance in digital assessments by examining whether spreadsheet-based items in PISA introduce additional difficulty unrelated to mathematical knowledge. Together, these analyses illustrate how a model-based, theory-informed approach can reveal both strengths and tensions in assessment design and contribute to more robust validity arguments for interpreting scores from large-scale mathematics assessments.

Session 5, 8.50-9.10

Applying a Rasch Tree Approach to Identify DIF in the Bullying Scale of PIRLS 2021 across 15 Countries

Purya Baghaei
IEA Hamburg, Germany

Introduction:
Large-scale assessments increasingly rely on background questionnaire scales to interpret student experiences, yet the linguistic and cultural equivalence of these scales remains uncertain. Ensuring measurement equivalence across diverse linguistic and cultural contexts is a central challenge for large-scale assessments. Background questionnaire scales, though non-cognitive in intent, may still be influenced by respondents' reading ability or sociocultural norms. This study examined differential item functioning (DIF) in the PIRLS 2021 Exposure to Bullying student scale across fifteen countries, focusing on the effects of reading achievement and gender on item functioning.
Methods:
The Partial Credit Tree (PCtree) method was applied separately within each country to detect DIF. Reading achievement and gender were used as covariates. Items with pγ categories B or C were identified as exhibiting moderate or large DIF. Item content, country-level reading performance, and tree structures were examined to interpret the nature and sources of DIF.
Results:
Analyses revealed systematic non-invariance across countries. Reading achievement emerged as the predominant splitting variable indicating that comprehension skill influenced responses to non-cognitive items. Gender appeared as a frequent secondary splitter, interacting with reading ability in several contexts. Reading-related DIF dominated cyberbullying items while gender-related DIF was more evident for physical and relational aggression items.
Discussion:
The findings demonstrate that background questionnaire items are sensitive to both linguistic complexity and sociocultural interpretation. Reading-related DIF corresponds

Session 5, 8.50-9.10

closely with national reading achievement levels, and gender DIF reflects differing norms of aggression and reporting. Incorporating readability checks, cultural adaptation, and data-driven DIF detection into item development and translation can substantially improve validity and cross-national comparability in large-scale educational assessments.

Session 5, 9.10-9.30

Applying Mixture IRT Models to a High-Stakes Reading Comprehension Test to Detect L2 Reader Profiles

Farshad Effatpanah
Islamic Azad University, Mashhad Branch, IR
Faculty of Rehabilitation Sciences, TU Dortmund University, DE

An important assumption of conventional item response theory (IRT) models is parameter invariance, indicating that item parameters remain constant across all examinees in the population. In fact, the ordering of item parameters should be the same for all examinees. However, this assumption may be violated when there are qualitative differences among examinees, such as variations in their cognitive strategies. Mixture IRT (MixIRT) models relax this assumption and allow item parameters to differ across latent classes such that each item can possess distinct parameter values within each class. They combine the IRT framework's capacity to model individual differences within latent classes with the latent class model's (LCM) ability to identify qualitatively distinct subgroups.
This study applied multiple MixIRT models to the reading comprehension section of a high-stakes multiple-choice language test to investigate heterogeneous profiles of second/foreign language (L2) readers. Latent classes were further characterized based on examinees' gender, lexico-grammatical knowledge, and overall language proficiency as measured by a Cloze test. Responses from 2,439 examinees to the reading section were analyzed using several MixIRT models, including the mixture Rasch model, 2-parameter logistic (2PL) MixIRT, 3PL MixIRT, and 4PL MixIRT models, each estimated with one to six latent classes. The 2PL IRT model with two latent classes provided the best model fit.
The two classes were: (1) Local Processors and (2) Global Integrators. Class 1 included lower- to intermediate-proficiency examinees with limited overall language ability and lexico-grammatical knowledge, who predominantly employed bottom-up, sentence-level processing and tended to rely on superficial strategies, including rote memorization of

Session 5, 9.10-9.30

isolated lexical and grammatical forms. However, Class 2 comprised higher-proficiency examinees who demonstrated stronger general language ability and lexico-grammatical knowledge and were able to coordinate both top-down and bottom-up processes. These readers integrated higher and lower-level (sub)skills and employed predictive and inferential strategies to construct coherent, text-level understanding.

Session 5, 9.30-9.50

The Psychometric Evaluation of the Strengths and Difficulties Questionnaire: What Does It Really Measure?

Bengü Börkan
Boğaziçi University, TR

The Strengths and Difficulties Questionnaire (SDQ) is a widely used tool for assessing adolescents' psychosocial adjustment, yet its underlying factor structure remains debated. This study examined the factorial validity and item-level measurement properties of the Turkish self-report SDQ (ages 11–17) using data from 1,458 middle and high school students. Confirmatory factor analysis (CFA) with the WLSMV estimator was used to evaluate several theoretically grounded models, including the original five-factor model, competing three-factor models, a five-factor model with a method factor, a second-order five-factor model, and a bifactor model.

The original five-factor model showed generally acceptable fit, with strong loadings for most items, although the Peer Problems subscale displayed weaker and inconsistent associations. The three-factor model performed worse, indicating reduced construct coherence. Incorporating a latent method factor for reversed-worded items improved fit without altering substantive factor structure. Second-order CFA with all five factors loading onto a general factor did not converge, whereas a modified second-order model—where four factors loaded onto a general difficulties factor and Prosocial Behavior covaried with it—showed moderate fit. The general difficulties factor was positively associated with conduct, emotional, and hyperactivity-inattention problems but negatively with peer problems, while Prosocial Behavior was negatively associated with overall difficulties.

A bifactor model provided the best fit, capturing both a general factor and specific factors simultaneously. Findings suggest that although a general difficulties factor exists, most SDQ

Session 5, 9.30-9.50

variance is explained by specific factors, supporting the hierarchical structure but highlighting limitations in the Peer Problems subscale. Across models, CFA results indicated suboptimal fit and persistent item-level misfit, particularly for peer problems and reverse-worded items. To further investigate, exploratory structural equation modeling (ESEM) was used to examine cross-loadings and residual associations, and Rasch modeling evaluated item fit, scale targeting, and reliability. Rasch results identified several misfitting items and limited discrimination within subscales, again most notably for peer problems.
Overall, this study supports the revision of the SDQ. The integration of CFA, ESEM, and Rasch analyses provides converging evidence of structural and measurement challenges within the SDQ and highlights the importance of combining factor-analytic and item-response approaches when evaluating adolescent self-report instruments.

SAMC Session 6

Session 6

Mattias Bohm
Jeanette Melin
David Andrich

Session 6, 10.30-10.50

Rasch analysis of the Zarit Burden Interview (ZBI-22) in family members of out-of-hospital cardiac arrest survivors

Mattias Bohm
Brain Injury After Cardiac Arrest Research Unit, Department of Clinical Sciences Lund, Lund University, SE
Department of Intensive and Perioperative Care, Skåne University Hospital, Malmö, SE

Introduction: Family members of out-of-hospital cardiac arrest (OHCA) survivors may

experience substantial caregiver burden, particularly when survivors have cognitive problems. The aim was to evaluat the measurement properties of the 22-item Zarit Burden Interview (ZBI-22) in family members of OHCA survivors.
Methods: Data were drawn from the cognitive substudy of the Target Temperature Management 33°C versus 36°C after Out-of-Hospital Cardiac Arrest trial (TTM trial). At 6 months, 271 family members of 287 survivors completed the ZBI-22. Measurement properties were assessed using a polytomous Rasch model in RUMM2030, examining overall fit, item and person fit, local independence, response category functioning, targeting, reliability, and differential item functioning (DIF).
Results: Of the 271 respondents, 220 (81%) were women; mean age was 55 years. Total item chi-square (217.0; df 66; $p < 0.001$) indicated insufficient overall model fit. Mean (SD) fit residuals were −0.42 (2.08) for items and −0.34 (1.11) for persons. Mean person location was −1.85 logits; 11 individuals (4%) had the minimum score and 10 (4%) had person fit residuals < −2.5. Seven items showed significant chi-square probabilities after Bonferroni adjustment, and four items had fit residuals < −2.5. Disordered thresholds were present in 14 items. Local dependency was suggested by residual correlations in 11 item pairs, based on Yen's Q3 critical values exceeding the mean residual correlation. The PCA/t-test protocol indicated potential multidimensionality, with 11.5% (95% CI: 8.9–14.2) of individuals showing significantly different person estimates. Uniform DIF was found for one item by sex and two items by age. Reliability was acceptable (Person Separation Index 0.86). Maximum information occurred at 0 logits, corresponding to a ZBI-22 total score of 26.
Discussion: The ZBI-22 showed acceptable reliability but suboptimal Rasch model fit, with signs of multidimensionality, local dependency, and response category problems. Respondents' lower burden relative to the centralised item mean indicates mistargeting. A score of 26 may be a useful cut-off for caregiver burden in this population. Future studies should evaluate revised response categories, test shorter ZBI versions, and use external anchors and cognitive interviews to support content validity.

Session 6, 10.50-11.10

The Older Persons' Well-Being Scale (OPWELLS): From a Reversed Construct Theory Approach to the Development of a Digital App

Jeanette Melin
Linneaus University, Faculty of Health and Life Science, Kalmar
Swedish Defence University
Department of Leadership, Demand and Control Karlstad
SE

Background: Recent studies describe inconsistency with the definitions and heterogeneity of items in existing well-being scales for older persons, and only a few of these scales have undergone psychometric evaluation within a formal measurement framework.
Aim: The research project aimed to enhance measurement quality for measuring older persons' well-being through the development of the Older Persons' Well-being Scale (OPWELLS).
Method: The development of the OPWELLS has followed a reverse-construct theory-building approach. Content specification and item development were based on established well-being and quality-of-life scales and informed by a thematic analysis of underlying dimensions and approaches to measuring positive mental health. Cognitive interviews (n=26) were conducted with older persons aged 65–102 years, using think-aloud and verbal probing techniques. Two stages of data collection and psychometric analyses were conducted according to the Rasch model. In the first stage (n=441), the focus was on identifying potentially redundant items, and in the second stage (n=1882), stepwise item removal and/or alternation were performed guided by considerations of both content representation and measurement properties. Subsequently, a Shiny app is also under development to transform raw scores into linear measures and facilitate visualization of group-level data.
Results: An item pool of 21 items covering key dimensions of well-being was developed and subject to cognitive interviews. Most participants got a positive experience, engaging easily with the items and reflecting meaningfully on their responses. The interviews also yielded feedback-informed targeted revisions, including adding a complementary item on personal relationships and reordering the most cognitively demanding items. In the first stage of

Session 6, 10.50-11.10

psychometric analyses, three items were found to be highly locally dependent on another similar item, and another item had an exceptionally deviating misfit. The second stage of psychometric analyses revealed 10 items with optimised content coverage and satisfactory measurement properties.
Conclusion: The OPWELLS, developed through a reverse-construct theory-building approach and evaluated using the Rasch model, provides a psychometrically sound and content-relevant scale for assessing well-being among older persons. The accompanying Shiny app further enhances the practical utility of OPWELLS by enabling the transformation of raw scores into linear measures and facilitating interpretation at the group level.

Session 6, 11.10-11.30

An application of Rasch's physical science methods to psychological science data: A quantitative, meta-metre characterisation of growth on the Stanford-Binet intelligence test

David Andrich
The University of Western Australia, AU

The British physicist Geoffrey West (2017) summarised the study of growth on physiological variables, for example weight and metabolic rate, from the approach of the physical sciences. This approach led to laws of physiological growth that are of the kind found in the physical sciences. West's approach to the study of a physiological variable from the perspective of the physical sciences was anticipated in the 1950s by the Danish Mathematician Georg Rasch, summarised in Rasch (1977) and Olsen (2003). This paper presents an adaptation of Rasch's approach to characterise the quantitative rate of growth on the Stanford-Binet intelligence test (SBIT) across a 12- year life span from the age of three in a manner of a physical law. It shows a very accurate characterisation, with over 98% of the variance accounted for, which also shows no gender differences in the rate of growth. In particular, the hypothesises that the rate of growth at any time of measurement is effectively characterised by a single factor which is inversely proportional to the square of the current measurement is confirmed.

References:
Rasch, G. (1977). On specific objectivity: an attempt at formalising the request for generality and validity of scientific statements. Danish Yearbook of Philosophy 14, 58-94.

Session 6, 11.10-11.30

West, G. (2017). Scale: The Universal Laws of Life, Growth and Death in Organisms, Cities and Companies. New York. Penguin Books.
Wolk Olsen, L. (2003). Essays on Georg Rasch and his contributions to statistics. University of Copenhagen, Institute of Economics. Copenhagen.

SAMC Session 7

Session 7

Mark Elliott
Svend Kreiner
Marianne Müller
Magnus Johansson

Session 7, 13.00-13.20

Quantifying the magnitude of item interaction effects in subtests exhibiting local item dependence

Mark Elliott
Department of Computer Science & Technology, University of Cambridge
Cambridge University Press & Assessment
UK

Approaches to investigating local item dependence (LID) in Rasch models are mostly based on tests for LID between pairs of items, such as Yen's Q3 (Yen, 1993), a drawback of which is that they may not be sensitive to LID effects that are small between individual item pairs but which act cumulatively across a set of items. Zenisky, Hambleton and Sireci (2002) developed an approach to cumulative LID involving creating a superitem based on the sum score of a set of items and comparing the overall test reliability of the resulting test with that of the original test.

We first explore some effects cumulative LID can have on item estimates, then propose an extension to Zenisky et al's method which involves comparing the thresholds of a sum-score superitem with those of a superitem created to have the same item response function (IRF) as the sum of the IRFs of the individual subtest items, derived mathematically. A compression of the scale of the thresholds of the former compared to the latter should indicate the presence of LID (Andrich, 1985).

We explore the utility and limitations of this approach via simulated data where the response to each item in a set directly affects the difficulty of the next item, under two different conditions: when the order of items is prescribed, and when the items may be taken in any order.

Session 7, 13.00-13.20

We find that our novel approach is sensitive to cumulative LID, in particular where the extent of LID between each pair of items is too small for pairwise approaches to detect effectively, and provides a means of attempting to quantify the magnitude of item interactions caused by LID.

References

Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. Test design: Developments in psychology and psychometrics, 245-275.
Yen, W. M. (1993). Scaling performance assessments: Strategies for managing LID. Journal of Educational Measurement, 30(3), 187-213.

Zenisky, A. L., Hambleton, R. K., & Sired, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. Journal of Educational Measurement, 39(4), 291-309.

Session 7, 13.20-13.40

Exact conditional tests of person fit to Rasch models and log-linear Rasch models

Svend Kreiner, University of Copenhagen, DK

specific population of persons and a specific response process, and Rasch' definition of specific objectivity underscore that measurement may not be objective outside the frame of reference. It is for this reason, that tests of person fit are essential for clinical single patient applications of Rasch measurement. Measurement may be confounded for the following four reasons out of which the first three raises concerns with the frame of

reference.

- First, because of aberrant response behavior.
- Second, because the patient is not a member of the population defined by the frame of reference.
- Third, because of incorrect scoring of responses.
- Finally, because of statistical Type I errors.

We summarize results on an exact conditional test of person fit in Rasch models proposed more than fifty years ago by the Swedish mathematician Per Martin Löf and extend them to log-linear Rasch models with local dependence among items. The test rejects person misfit if the person's pattern of the responses to items is improbable according to the conditional distribution of the response pattern given the total score over all items.

The exact conditional test of person fit is in many ways similar to standardized log-likelihood Lz test of person fit in IRT models that assumes that items are locally independent. Lz rejects person fit if the observed response pattern if the IRT model claims that the patterns is improbable if the true value of the person parameter is equal to the estimate of the person parameter. The presentation will elaborate on the similarities between the two tests and show that the exact conditional test of person fit in Rasch models sidesteps a number of problems that impede the Lz because of the bias and error of estimates of person

Session 7, 13.40-14.00

Criterion-referenced interpretation of quantitative measurement of the amount of medication issues by DISABKIDS

Marianne Müller, Bern University of Applied Sciences, CH

Measurement by Rasch models is quantitative, interval-scaled, sample-free, and – given the correct type of estimators – specific objective. Quantitative measurement is useful for statistical applications studying associations between PROMS and other health related variables and it may be useful for ranking of persons. However, scores and estimates of person parameters rarely provide information that is useful in connection with classroom applications of formative testing or in connection with clinical single patient applications of health related scales physicians need concrete qualitative information on health related issues.

Criterion-referenced interpretability of test scores is the degree to which one can assign qualitative meaning to an instrument's scores or change of scores, which is sample-independent and does not depend on other persons' scores. Interpretability is not considered to be a measurement property, but it is an important requirement for intelligent use of a measurement instrument.

Kreiner et al. (2025) describe criterion-referenced interpretation of educational test results by systematic analyses of scale-anchored probabilities of responses to items from different item-domains. This presentation will elaborate on the methodology and illustrate how criterion-referenced interpretation of measurement by DISABKIDS may provide concrete qualitative information on the health related quality of life issues of children with Type 1 diabetes.

Reference

Svend Kreiner, Marianne Müller & Tine Nielsen (2025) Criterion-referenced interpretation of educational test results: illustrated by analysis of Danish results from PIRLS 2016. Submitted to Educational Methods and Psychometrics.

Session 7, 14.00-14.20

Assessing model and item fit in psychometrics – moving toward up-to-date methods and appropriate cutoff values

Magnus Johansson, Karolinska Institutet, SE

Psychometric analyses rely on the interpretation of various fit metrics to evaluate model and item fit to the data. Critical values for these metrics, also known as cutoff values, are usually applied based on rule-of-thumb, sometimes referring to simulation studies such as the ubiquitous Hu & Bentler (1999), books like Bond & Fox (2015), or seemingly sage advice from field dignitaries. However, in both confirmatory factor analysis (CFA) and Rasch measurement theory (RMT), it has long been known that frequently used rule-of-thumb cutoffs are inappropriate for general use. With the computational power of modern laptops, it is now feasible and desirable to use local simulations to determine customized and appropriate cutoff values for various fit metrics for the dataset being analyzed.

This presentation will describe and demonstrate freely available simulation methods for several central aspects of model and item fit assessment primarily within RMT, with some examples pertaining to CFA. Key simulation studies will be summarized, and recommendations will be made regarding both analytical methods to leave behind and what to use instead.