The 10th European Congress of Methodology (EAM2023)

Book of Abstracts

Version of 3 July 2023

Contents

1 Tuesday 11 July **4**

1.1 Keynote speaker 10h00–11h00 . . . 5
1.2 State-of-the-art 15h00–15h30 Aud 1 . . . 6
1.3 State-of-the-art 15h00–15h30 Aud 2 . . . 7
1.4 Parallel sessions 11h30–13h00 Auditorium 1 . . . 8
1.5 Parallel sessions 11h30–13h00 Auditorium 2 . . . 13
1.6 Parallel sessions 11h30–13h00 Auditorium 3 . . . 18
1.7 Parallel sessions 11h30–13h00 Auditorium 4 . . . 22
1.8 Parallel sessions 11h30–13h00 Lecture room 1.2 . . . 26
1.9 Parallel sessions 11h30–13h00 Lecture room 1.3 . . . 30
1.10 Parallel sessions 16h00–17h30 Auditorium 1 . . . 34
1.11 Parallel sessions 16h00–17h30 Auditorium 2 . . . 39
1.12 Parallel sessions 16h00–17h30 Auditorium 3 . . . 44
1.13 Parallel sessions 16h00–17h30 Auditorium 4 . . . 49
1.14 Parallel sessions 16h00–17h30 Lecture room 1.2 . . . 53
1.15 Parallel sessions 16h00–17h30 Lecture room 1.3 . . . 57
1.16 Poster session 1 14h00–15h00 . . . 61
1.17 Panel discussion 17h30–18h30 . . . 91

2 Wednesday 12 July **92**

2.1 Keynote speaker 10h00–11h00 . . . 93
2.2 State-of-the-art 15h00–15h30 Aud 1 . . . 94
2.3 State-of-the-art 15h00–15h30 Aud 2 . . . 95
2.4 Parallel sessions 08h30–10h00 Auditorium 1 . . . 96
2.5 Parallel sessions 08h30–10h00 Auditorium 2 . . . 101
2.6 Parallel sessions 08h30–10h00 Auditorium 3 . . . 106
2.7 Parallel sessions 08h30–10h00 Auditorium 4 . . . 111
2.8 Parallel sessions 08h30–10h00 Lecture room 1.2 . . . 115
2.9 Parallel sessions 08h30–10h00 Lecture room 1.3 . . . 119
2.10 Parallel sessions 1h30–13h00 Auditorium 1 . . . 123
2.11 Parallel sessions 1h30–13h00 Auditorium 2 . . . 128
2.12 Parallel sessions 1h30–13h00 Auditorium 3 . . . 133
2.13 Parallel sessions 1h30–13h00 Auditorium 4 . . . 137
2.14 Parallel sessions 1h30–13h00 Lecture room 1.2 . . . 141
2.15 Parallel sessions 1h30–13h00 Lecture room 1.3 . . . 145
2.16 Parallel sessions 16h00–17h30 Auditorium 1 . . . 149
2.17 Parallel sessions 16h00–17h30 Auditorium 2 . . . 154
2.18 Parallel sessions 16h00–17h30 Auditorium 3 . . . 159
2.19 Parallel sessions 16h00–17h30 Auditorium 4 . . . 164
2.20 Parallel sessions 16h00–17h30 Lecture room 1.2 . . . 168
2.21 Parallel sessions 16h00–17h30 Lecture room 1.3 . . . 171
2.22 Poster session 2 14h00–15h00 . . . 175

3 Thursday 13 July **208**
3.1 Keynote speaker 10h00–11h00 . . . 209
3.2 State-of-the-art 15h00–15h30 Aud 1 . . . 210
3.3 State-of-the-art 15h00–15h30 Aud 2 . . . 211
3.4 Parallel sessions 08h30–10h00 Auditorium 1 . . . 212
3.5 Parallel sessions 08h30–10h00 Auditorium 2 . . . 216
3.6 Parallel sessions 08h30–10h00 Auditorium 3 . . . 221
3.7 Parallel sessions 08h30–10h00 Auditorium 4 . . . 224
3.8 Parallel sessions 08h30–10h00 Lecture room 1.2 . . . 228
3.9 Parallel sessions 08h30–10h00 Lecture room 1.3 . . . 232
3.10 Parallel sessions 11h30–13h00 Auditorium 1 . . . 236
3.11 Parallel sessions 11h30–13h00 Auditorium 2 . . . 241
3.12 Parallel sessions 11h30–13h00 Auditorium 3 . . . 245
3.13 Parallel sessions 11h30–13h00 Auditorium 4 . . . 249
3.14 Parallel sessions 11h30–13h00 Lecture room 1.2 . . . 253
3.15 Parallel sessions 11h30–13h00 Lecture room 1.3 . . . 257
3.16 Poster session 3 14h00–15h00 . . . 261

1 Tuesday 11 July

1.1 Keynote speaker 10h00–11h00

Title

Bringing owls to Athens: Some thoughts on what makes a good simulation study in methodological research

Author

Carolin Strobl

University of Zurich

Abstract

Simulation studies are an important tool for investigating and comparing quantitative methods with respect to various performance indicators. Many of us use and read about simulation studies on a daily basis in their methodological research. Just like the empirical sciences, who have been forced by the replication crisis to debate and reflect upon their accustomed tools for accumulating knowledge, I believe our discipline would profit if every now and then we thought and talked about why and how we typically do simulation studies. In this spirit, I dare to bring owls to Athens by presenting some thoughts on what makes a good simulation study, what distinguishes simulation studies from methods comparisons based on empirical data sets, and whether/how open science practices like preregistration can improve our ways of accumulating knowledge on quantitative methods like they do for research questions in the empirical sciences.

1.2 State-of-the-art 15h00–15h30 Aud 1

Title

Single-case experimental designs (and) data analysis: One size does not fit all

Author

Rumen Manolov

University of Barcelona

Abstract

Single-case experimental designs (SCEDs) allow obtaining empirical evidence regarding what intervention works for whom, under what circumstances. According to whether the target behavior is reversible or not and according to the degree to which immediate intervention effects are expected, different SCEDs can be used: multiple-baseline, reversal, alternating treatments, changing criterion, as well as hybrid designs involving combinations of the four main types.

The analysis of SCED data entails several challenges, such as: (a) the presence of autocorrelation in the data; (b) multiple data features on which the analysis can focus: level, trend, variability, immediacy, overlap, and consistency; (c) increasing number of data analytical options (more than twenty), some of which remain largely untested, whereas others have been object of multiple studies whose results are not easily summarized in a couple of sentences; (d) the divide between the typical data analytic approach of applied researchers (used to analyzing data visually with none or few visual aids and preferring simple quantifications such as nonoverlap indices) and the statistical developments in the field (such as multilevel modeling, standardized mean difference taking autocorrelation into account, or randomization tests); (e) the distinction between formative and summative analysis, and its relation to the possibility of preregistration; and (f) the absence of most of the data analytical options from typical software such as SPSS.

In this ocean of challenges and possibilities, applied researchers, however, are not alone. There are reporting guidelines, methodology quality appraisal tools (rubrics), articles illustrating the use of different analytical techniques, articles providing suggestions that can help choosing among data analytic techniques, and freely available user-friendly websites implementing many data analytical options and graphically representing the data. The current talk, and slides that accompany it, provide links to many of these resources.

1.3 State-of-the-art 15h00–15h30 Aud 2

Title

Theory-based hypothesis evaluation using information criteria for one and multiple studies

Author

Rebecca Kuiper

Utrecht University

Abstract

The evaluation of hypotheses is a key feature of research in the behavioral, social, and biomedical sciences. For example, researchers might be interested in whether medicine A works better (e.g., leads to more happiness) than medicine B, which works better than a placebo (in an ANOVA model: $\mu_A > \mu_B > \mu_{Placebo}$); or: number of children is a stronger predictor for happiness than income and age (in a regression model with standardized parameters: $\beta_{NoC} > \{\beta_{Inc}, \beta_{age}\}$); or: the cross-lagged effect of stress to anxiety, of rumination to stress, and of rumination to anxiety are higher than the cross-lagged counterparts (in a random-intercept cross-lagged panel model: $\phi_{SA} > \phi_{AS}, \phi_{RS} > \phi_{SR}, \phi_{RA} > \phi_{AR}$). Evidence for such **informative, theory-based hypotheses** can be obtained via (information-theoretical and Bayesian) model selection.

In my presentation, I will focus on information-theoretical model selection, that is, model selection using information criteria. I will introduce the AIC-type criterion GORIC and its approximation the GORICA. I will discuss how the GORIC(A) results can be interpreted and how they do quantify the support for the hypothesis (as opposed to null hypothesis testing). I will address two cases: i) the case when there is one study and ii) the meta-analytic case when there are multiple studies from which we would like to aggregate the results.

1.4 Parallel sessions 11h30–13h00 Auditorium 1

Symposium Overview

Detecting and understanding aberrant response behaviors—Methodological advances and applications

Author(s)

Esther Ulitzsch[1,2], Johan Braeken[3]

[1]IPN - Leibniz Institute for Science and Mathematics Education, Kiel, Germany; [2]University of Mannheim, Mannheim, Germany; [3]CEMO: Centre for Educational Measurement at the University of Oslo, Oslo, Norway

Abstract

Research in the social sciences heavily relies on questionnaire data. Aberrant response behaviors, such as disengaged responding in the context of low-stakes questionnaire administrations, faking in high-stakes contexts, or excessive omission behavior, pose a major threat to the validity of conclusions drawn from questionnaire data. Understanding their occurrence is important for gauging data quality and may ultimately inform questionnaire designs that curb aberrant response behaviors. This symposium highlights methodological advances and applications in studying the occurrence of different aspects of aberrant response behaviors and their underlying mechanisms in a broad array of research contexts. The first study investigates faking behavior and applies count and log-normal item response models to process data such as response times or answer changes to study how different stakes, person and item characteristics manifest in response editing. The second study employs cross-classified mixed effects models to study gender differences in item nonresponse patterns in large-scale assessment background questionnaires across countries, scales' formats, and contents. The third study introduces a mixture item response theory model to classify and flag potential random responders and illustrates its utility by studying the prevalence of random responding on a scale as a function of progression through the questionnaire. The fourth study presents a mixture modeling approach that leverages time spent on screen for investigating the trajectory of careless and insufficient effort responding in ecological momentary assessment data.

Title

Random Responders on the TIMSS 2015 student questionnaire

Author(s)

Johan Braeken, Saskia Van Laar

CEMO: Centre for Educational Measurement at the University of Oslo, Oslo, Norway

Abstract

The typical low-stakes character of student questionnaires in educational research often raises doubts about whether all students fully engage with the questionnaire and not just randomly fill in their responses. In the absence of any auxiliary log data, only the atypicality of the pattern of responses given on the different scales in the questionnaire can be somewhat informative to detect such random responders. Following a model expansion approach to assess measurement appropriateness, we use a mixture item response theory (IRT) model to classify and flag potential random responders. Results on the prevalence of random responders on a scale as a function of progression through the questionnaire (including factors such as questionnaire length, scale position, and contents) are presented for the student questionnaire of the Trends in Mathematics and Science Study (TIMSS) 2015.

Symposium title

Detecting and understanding aberrant response behaviors—Methodological advances and applications

Title

A screen-time-based mixture modeling approach for detecting and investigating careless and insufficient effort responding in ecological momentary assessment data

Author(s)

Esther Ulitzsch[1,2], Gabriel Nagy[1], Oliver Lüdtke[1], Steffen Nestler[3]

[1]IPN - Leibniz Institute for Science and Mathematics Education, Kiel, Germany; [2]University of Mannheim, Mannheim, Germany; [3]University of Münster, Münster, Germany

Abstract

Ecological momentary assessment (EMA) involves repeated real-time sampling of respondents' current behaviors and experiences. The intensive repeated assessment imposes an increased burden on respondents, rendering EMA vulnerable to respondent non-compliance and/or careless and insufficient effort responding (C/IER). We developed a mixture modeling approach that equips researchers with a tool for (a) gauging the degree of C/IER contamination of their EMA data, (b) studying the trajectory of C/IER across the study, and (c) investigating occasion and person characteristics associated with increased C/IER rates. For separating attentive from C/IER behavior, the approach leverages collateral information from screen times, which are routinely recorded in smartphone-administered EMAs, and translates theoretical considerations on respondents' behavior into component models for attentive and careless screen times as well as for the functional form of C/IER trajectories. We show how a sensible choice of components models (a) allows disentangling short screen times due to C/IER from familiarity effects due to repeated exposure to the same measures and (b) aids in gaining a fine-grained understanding of C/IER trajectories by distinguishing within-day from between-day effects. The approach is illustrated on EMA data from the German Socio-Economic Panel innovation sample.

Symposium title

Detecting and understanding aberrant response behaviors—Methodological advances and applications

Title

Gender Differences in Item Nonresponse in the PISA 2018 Student Questionnaire

Author(s)

Kseniia Marcq, Johan Braeken

CEMO: Centre for Educational Measurement at the University of Oslo, Oslo, Norway

Abstract

Gender differences in item nonresponse are well-documented in high-stakes achievement tests, where female students are shown to omit more items than male students. These gender differences in item nonresponse are often linked to differential risk-taking strategies, with females being risk-averse and unwilling to guess on an item, even if it could gain them credits. In low-stakes settings, similar trends should not apply, as the students carry no consequence for their performance. Instead, test-taking motivation is argued to be the pivoting factor, with female students seen as more motivated and omitting fewer items than male students. In contrast to the high- and low-stakes achievement tests, less is known about gender differences in item nonresponse in student background questionnaires.Using cross-classified mixed effects models, we examined gender differences in item nonresponse on the Programme for International Student Assessment (PISA) 2018 student questionnaire across 80 countries and 71 scales. On average, the odds of male students omitting a questionnaire item were double the odds of female students, consistent with the expected trend in the low-stakes setting. However, we show that gender differences in item nonresponse are not merely a function of stakes involved but a more complex phenomenon that is context-dependent and not necessarily stable across countries, scales' formats, and contents. We argue that examining differences in item nonresponse patterns could serve as a source of additional information about the students' test-taking behavior and the quality of the questionnaire.

Symposium title

Detecting and understanding aberrant response behaviors—Methodological advances and applications

Title

Using Process Data to Understand Response Processes Underlying Socially Desirable Responding in Forced-Choice Questionnaires

Author(s)

Susanne Frick[1], Anna Brown[2]

[1]TU Dortmund University, Dortmund, Germany; [2]University of Kent, Canterbury, United Kingdom

Abstract

Impression management (aka Faking) on self-report questionnaires is a concern in high-stakes assessments. The forced-choice (FC) format has been proposed to overcome faking. However, faking resistence depends on the item desirability and on the desirability matching. Process data, such as response times or changes made to initial response, can help understand the process of faking. The objective of this study is to to investigate how item and person characteristics related to socially desirable responding manifest in response editing in FC questionnaires.

In this study, participants responded to 69 item blocks in a forced-choice format presented on cards with a drag-and-drop display. We tracked the clicks on the cards and the mouse moves in and out of the cards, together with the corresponding timestamps, and the actual card selections. Item desirabilities were rated on a rating scale by a separate sample.

We employed count item response models to the numbers of moves and clicks and log-normal item response models to the response times. Items' social desirability valence and ambiguity served as item-level covariates. Faking ability was measured by a Thurstonian IRT model fitted to whether the optimal rank order as defined by the mean social desirability ratings, was chosen.

From a psychometric perspective, this study can inform further psychometric developments for the analysing of process data. From a practical perspective, the results of this research can inform the development of fake-resistant assessments and facilitate the evaluation of the impact of faking on current assessments.

Symposium title

Detecting and understanding aberrant response behaviors—Methodological advances and applications

1.5 Parallel sessions 11h30–13h00 Auditorium 2

Symposium Overview

Large-Scale, Sparse and Noisy Educational Data: Challenges, Solutions, and Some Exemplary Findings

Author(s)

Martin J. Tomasik

University of Zurich, Zurich, Switzerland

Abstract

This symposium brings together a group of researchers working with and trying to make sense of big data from a computer-based formative feedback system called MINDSTEPS. This system has been running for some years now and comprises data of 100K+ students from primary and secondary schools in four cantons of Switzerland. MINDSTEPS comprises an itembank with 10K+ items in eleven distinct subject domains (such as "German: Reading Compehension", "German: Grammar", "English: Grammar", or "Mathematics: Numbers and Variables") that allows both teachers and students putting together assessments for learning, practising, and providing/obtaining formative feedback.

Assessment situations tend to vary heavily from teacher to teacher, student to student, and time point to time point. From a methodological perspective, this situational heterogeneity results in a low signal-to-noise ratio and makes the modelling of learning trajectories and the detecting of effects quite challenging. Another challenge of this data set is its sheer size and sparsity. Although there are tens of thousands of items and hundreds of thousands of students measured at multiple time points, of course not every student has repeated measurement on all items.

Against this backdrop, we present solutions that we developed to work with the data set along with some exemplary findings. More specifically, we address different variants of vertical scaling approaches, ask about methods for detecting item misfit, describe long-term learning trajectories of the students and heterogeneity in these trajectories, report findings on the dynamic interrelations between scales, and deal with the issue of multi-dimensionality in the data set.

Title

Detecting parameter instability in large assessments: An adaptation of score-based tests

Author(s)

Rudolf Debelak, Charles C. Driver

University of Zurich, Zurich, Switzerland

Abstract

An important step in the psychometric analysis of educational scales is the investigating measurement invariance. In the context of large scale assessments such as MINDSTEPS, traditional approaches for checking measurement invariance can be computationally demanding and even infeasible. Furthermore, the psychometric evaluation of educational assessments requires the detection of violations of parameter invariance with regard to categorical, ordinal and continuous covariates such as gender, age or educational level. We present an adaptation of score-based tests for checking measurement invariance in the framework of item response theory as a possible solution. As a first contribution, we outline a novel adaptation of these tests to frequentist and Bayesian estimation methods in large scale assessments. As a second contribution, we evaluate the new method using simulation studies. We find the new method to be sensitive to violations of measurement invariance while having a low Type I error rate in sufficiently large samples. We briefly discuss possible applications of the new method in the context of large scale educational assessments.

Symposium title

Large-Scale, Sparse and Noisy Educational Data: Challenges, Solutions, and Some Exemplary Findings

Title

Continuous-Time Modelling of Long and Short Term Relations Between Learning Domains

Author(s)

Charles Driver, Martin Tomasik

University of Zurich, Zurich, Switzerland

Abstract

We demonstrate how broad developmental theories may be instantiated as statistical models, using hierarchical continuous-time dynamic systems. This approach offers a flexible specification, and an often more direct link between theory and model parameters than common modelling frameworks, because direct relations between processes can be modelled – rather than modelling relations over some specific time interval as is typical of many approaches. The developmental theories in focus regard the relation between the academic competencies of mathematics and language, and we use data from the online learning system Mindsteps. Substantial consideration is given to parameter heterogeneity with respect to age, and we find measurements becoming more precise while correlations between scales reducing with age. In general however we find large positive correlation even between short-term fluctuations, which raises new questions. Modelling approaches such as this offer the potential for substantial insights into the relations between ongoing processes, but at substantial peril of discovering false insights due to model and interpretation issues, which we also devote time to.

Symposium title

Large-Scale, Sparse and Noisy Educational Data: Challenges, Solutions, and Some Exemplary Findings

Title

Vertical Scaling of a Huge but Sparse Data Set: Proprocessing Decisions, Modeling Variants, and the Challenge of Handling it All

Author(s)

Martin Tomasik[1], Charles Driver[1], Stella Bollmann[1], Benjamin Wolf[2]

[1]University of Zurich, Zurich, Switzerland; [2]Institute for Educational Evaluation, Zurich, Switzerland

Abstract

Vertical scaling maps ability measures that may have little to no overlap in their difficulty onto one single latent dimension. In the educational context, vertical scaling is needed to describe long-term learning growth of students across different school years, since items given to students in the earliest grades may be completely different to those given to the oldest. Given the structure of the MINDSTEPS data, we face three challenges when trying to implement vertical scaling. First, it is unclear how decisions concerning data preprocessing and modeling affect the estimation of item parameters and student abilities. Then, we need to find out how modelling decisions affect the robustness of parameter estimates. Finally, there are computational challenges of handling the large but sparse data set not allowing us to use exisisting software packages for this purpose. We present the results of model testing and the development of new calibration approaches, implemented in a newly developed R package capable of handling the large amounts of raw data at hand. Several models (i.e., from 1-PL to 4-PL, including or not including person-level and item-level covariates such as age, gender, type of item, time of assessment) have been compared with regard to their out-of-sample prediction performance and to grade-to-grade growth. Furthermore, we compared whether data collected in different assessments situations yielded different scaling results. Also, we systematically inspected the effects of different pre-processing decisions such as the minimum number of data points included or the consideration of unfinished assessments for parameter estimation.

Symposium title

Large-Scale, Sparse and Noisy Educational Data: Challenges, Solutions, and Some Exemplary Findings

Title

Detecting structure in data from a large-scale computer-based educational assessment system

Author(s)

Benjamín Garzón, Martin J. Tomasik

University of Zurich, Zurich, Switzerland

Abstract

A standard view in educational research posits that mathematical and verbal abilities are correlated but separate unidimensional, domain-specific factors. This perspective stands in contrast with the view from intelligence research that a general cognitive ability underlies the different academic abilities. The MINDSTEPS computer-based educational assessment system provides a large-scale, real-world dataset that can be used to investigate the latent structure of academic abilities over the school years in low-stakes scenarios, with higher ecological validity than standardized assessments. Here, we aimed to reveal the factorial structure of latent abilities from millions of student responses with an exploratory approach. The large size and high sparsity of the dataset pose a challenge for conventional multidimensional item-response theory techniques, and we therefore resorted to a regularized joint maximum likelihood method developed using a state-of-the-art machine learning framework to fit large-scale datasets. We found that only a few dimensions were sufficient to explain variation in latent abilities. One main factor accounted for most but not all of the variance in performance, was associated with age and displayed sexual dimorphism in its developmental trajectory, while the remaining factors did not show the same age dependence. The estimated set of factors, obtained from response data only, reflected the competencies in the school curriculum. Finally, we discuss the limitations of the method and the challenges and opportunities of the dataset for investigating related questions.

Symposium title

Large-Scale, Sparse and Noisy Educational Data: Challenges, Solutions, and Some Exemplary Findings

1.6 Parallel sessions 11h30–13h00 Auditorium 3

Title

Measuring the Dark Core of personality in Spanish speaking countries: Psychometric properties of the D-70 scale.

Author(s)

Jaime García-Fernández[1], Álvaro Postigo[1], Marcelino Cuesta[1], Covadonga González-Nuevo[1], Morten Moshagen[2]

[1]University of Oviedo, Oviedo, Spain; [2]University of Ulm, Ulm, Germany

Abstract

Research on socially aversive personality traits has shown that "dark" traits share many features among them. This fact led Moshagen et al. (2018) to conceptualize a common core (D), defined as a general tendency to maximize one's individual utility accompanied by beliefs that serve as justifications. This core can be assessed by the D-70 scale, which follows a bifactor structure. In this research, the Spanish version of the scale is validated, demonstrating its measurement invariance across sex and country based on a total of 12,356 participants from six countries (Spain, Mexico, Argentina, Colombia, Chile, Peru). Item analysis, reliability, and confirmatory factor analysis were performed for each specific country, following the measurement invariance analysis. Results showed that the Spanish version of the D70 is an essentially unidimensional and reliable measure and further supported invariance across sex and countries. These results could be useful for applied researchers who want to assess socially aversive personalities from the dark factor model, and for researchers interested in studying cultural or sexual differences on D in Spanish-speaking countries.

Oral presentations session title:

Measurement and Assessment

Title

Enhancing the methodological quality of intervention programs: Validity evidence of the Methodological Quality Scale

Author(s)

Salvador Chacón-Moscoso[1,2], Susana Sanduvete-Chaves[1], José A. Lozano-Lozano[2], F. Pablo Holgado-Tello[3]

[1]Universidad de Sevilla, Seville, Spain; [2]Universidad Autónoma de Chile, Santiago, Chile; [3]Universidad Nacional de Educación a Distancia, Madrid, Spain

Abstract

Professionals lack clear criteria for assessing the methodological quality (MQ) of intervention programs due to the wide variety of approaches to measuring MQ. Additionally, few studies present evidence of their metric properties or explain the reasoning behind the selected indicators. This work proposes the Methodological Quality Scale (MQS) as a simple and useful tool with reliability, validity evidence, and adequate metric properties. After a systematic review of the literature, two coders independently applied MQS to a set of primary studies on intervention programs at organizations. To obtain the validity facets in the scale, the number of dimensions was determined in parallel analyses before conducting factor analyses to identify the main dimensions. For each validity facet obtained, we presented basic descriptive statistics (mean, standard deviation, reliability, and mean discrimination). Furthermore, the validity facets scores were subjected to a theoretical interpretation based on Shadish, Cook and Campbell's validity model. The results included (a) an empirical validation of facets on MQ from a theoretical framework; and (b) an interpretation of the scores. The 10-item MQS specifies the inclusion criteria of the items, is easy to apply, and allows MQ profiles to be obtained in different areas of interventions.
Funding: This research was supported by the grant PID2020-115486GB-I00 funded by the Ministry of Science and Innovation –Ministerio de Ciencia e Innovación–, MCIN/AEI/ 10.13039/501100011 Government of Spain, Spain.

Oral presentations session title:

Measurement and Assessment

Title

The Behavioral, Emotional, and Social Skills Inventory (BESSI): Adaptation and Validation of a Spanish Version

Author(s)

Álvaro Postigo[1,2], Covadonga González-Nuevo[1], Jaime García-Fernández[1], Christopher J. Soto[3], Brent Roberts[4,5], Christopher M. Napolitano[4,6], Marcelino Cuesta[1]

[1]University of Oviedo, Oviedo, Spain; [2]University of Granada, Granada, Spain; [3]Colby College, Waterville, USA; [4]University of Illinois at Urbana-Champaign, Illinois, USA; [5]Tübingen University, Tubinga, Germany; [6]University of Zurich, Zúrich, Switzerland

Abstract

Social, emotional and behavioral skills comprise a broad set of abilities essential for establishing and maintaining relationships, regulating emotions, selecting and pursuing goals or exploring new stimuli. These have shown an important impact on people's lives. To improve their assessment in Spain, the aim of the present study was to adapt and validate the Behavioral, Emotional and Social Skills Inventory (BESSI) in the general Spanish population. The BESSI measures 32 facets of 5 domains through 192 items. Using two samples, one of the adolescents (12-17 years) and the other of the adult population (18 years and older), the psychometric properties of the Spanish version of the BESSI were studied (internal consistency, test-retest reliability, evidence of validity in terms of internal structure and in relation to other variables). The results show that, in general terms, each of the facets is unidimensional, with adequate internal consistency and stability of its scores and with adequate evidence of convergent validity in relation to the Big Five. In turn, the structure at the domain level is very similar to that of the original English version. The Spanish version of the BESSI has shown adequate psychometric properties, so it can be used in the general Spanish population to assess social, emotional and behavioral competencies, either in the adolescent or adult population.

Oral presentations session title:

Measurement and Assessment

Title

Modeling response options' position effects: Multi-group confirmatory factor analyses of the abstract reasoning test data

Author(s)

Vesna Buško

Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia

Abstract

A number of different sources of construct irrelevant variance has been recognized to impact on the test scores in performance-based measures. This study aimed to examine the effects of the position of the correct answer among the response options on the metric quality of the multiple-choice test items and the composite test scores. The analyses are based on the empirical data collected within university admission testing procedures regularly performed at the Faculty of Humanities and Social Sciences of the University of Zagreb. The presentation will focus on the test score data obtained on two of the three 20 item abstract reasoning subtests administered on the sample of 3234 applicants. Every applicant completed one out of four test forms differing only in the position of the correct response option on each item. Confirmatory factor models were specified so to include additional latent variables intended to account for the anticipated response option position variance. Multi-group analyses with groups defined by test forms taken confirmed the existence of the option position method factors in models specified for both subtests. Small but significant response option position variance was generally observed while the differences in meaning of the method factors identified for the two subtests could be attributed to the differences in the nature of tasks and the response mode referring to each subtest.

Oral presentations session title:

Measurement and Assessment

1.7 Parallel sessions 11h30–13h00 Auditorium 4

Title

(DE)BIASING?: The effect of straight-lining controls on assessments of risk attitudes and behaviours

Author(s)

Melisa Muric[1,2], Peter Thijssen[1], Tanja Perko[2,1], Catrinel Turcanu[2]

[1]University of Antwerp, Antwerp, Belgium; [2]Belgian Nuclear Research Center (SCK CEN), Mol, Belgium

Abstract

It is widely established that there are instances whereby survey respondents give inaccurate or false answers, and that response biases can affect the accuracy of data. With the rise of online surveys, and use of panel respondents, the risk of straight-lining increased. Survey companies reacted by building-in quality control features to inspect straight-lining, including; duration measures of the response process, "check questions" to verify attentiveness, and removal of straight-liners above a certain benchmark (e.g. 30%). Excluding straight-liners can however lead to new bias and loss of valid responses. We argue this especially when content-relevant differences between respondents are not considered.
Using survey data from 11 European countries on public attitudes and behaviors related to radon radiation, we evaluate, through SEM, the impact of the straight-lining controls of 5 different survey companies. To what extend do different straight-lining controls effectively reduce response style bias? We compare controls of survey companies to a scale specific method that disentangles content from style in balanced scales. Furthermore, we evaluate whether straight-lining controls introduce new bias. Respondents with less knowledge of radiation risk or experience with protective behaviors are likely more prone to straight-lining, due to lower engagement with- or higher perceived complexity of- the questionnaire. By systematically excluding them at a higher rate, the content-relevant composition becomes biased. There are also socio-demographical differences in straight-lining, however on these characteristics there are usually quota, so the sample composition is less skewed. Based on these findings we give methodological guidance for socio-technical research design and data-cleaning.

Oral presentations session title:

Response Styles

Title

Do people employ different response styles in their response strategies? An investigation with a mixture IRTree approach

Author(s)

Ömer Emre Can Alagöz, Thorsten Meiser

University of Mannheim, Mannheim, Germany

Abstract

In self-report studies, controlling for response style (RS) effects can improve the validity of our statistical inferences, which can be achieved with IRTree models. The traditional IRTrees model responses as an outcome of several distinct decision processes, namely informative response (whether to choose mid-category), direction (whether to choose dis/agreement categories), and intensity (whether to choose extreme categories). The substantive trait affects the decision about the direction, whereas midpoint RS (MRS) and extreme RS (ERS) largely determine the informative response and intensity decisions, respectively. Two limitations of traditional IRTrees are: 1) effects of the substantive trait on informative response and response intensity decisions are not accommodated, 2) all respondents are assumed to use ERS and MRS in their response strategy. However, respondents can differ in what types of RS they use and how strongly they use them. We address these limitations by proposing a mixture multidimensional IRTree (MM-IRTree) model that detects heterogeneity in response strategies, and it consists of four latent classes. A common substantive trait affects all decisions in all latent classes, but the classes differ in the set of RS that are used in response strategies. More specifically, class-specific response strategies include following RS: 1) ERS, 2) MRS, 3) both ERS and MRS, 4) neither ERS nor MRS. A simulation study showed excellent recovery of latent classes and reliable estimation of item/person parameters. An empirical analysis uncovered distinct classes of noticeable sizes suggesting that respondents indeed utilize different combinations of response styles.

Oral presentations session title:

Response Styles

Title

Scrutinizing response tendency to understand the cognitive processes: How can IRT help?

Author(s)

M.Carmen Navarro-González[1], José-Luis Padilla[1], Luis-Manuel Lozano[1], Álvaro Postigo[1,2]

[1]University of Granada, Granada, Spain; [2]University of Oviedo, Oviedo, Spain

Abstract

IR-Tree models assume that categorical item responses can best be explained by multiple response processes that take place in the judgement phase of item responding (Böckenholt, 2012, 2017): respondents, first, decide whether they agree or disagree to the item, and then how strong their agreement or disagreement is. The IR-Tree approach allows researchers examining response processes and detecting possible response styles (e.g., acquiescence, disaquiescence, extreme responding), disentangling response tendencies from the substantive trait measures. Following this approach, the aim of this study is to analyze multiple response processes and detect response styles of 11599 Spanish adolescents when responding to the "Sense of Belonging to School Scale" (SBSS) from the PISA 2018 Student Questionnaire (OECD, 2018). The SBSS consists of 6 four-point Likert-type items. We proposed three pseudo-items representing the nodes were created for each item: a) Agreement – Disagreement; b) Extreme Response (Disagreement) – Mild Response (Disagreement); and c) Extreme Response (Agreement) – Mild Response (Agreement). Once the tree was developed and the database was adapted, diverse R packages were used to analyze response processes of adolescents. Two IR-Tree models were tested: a) a descriptive model to detect extreme response styles, and b) an explanatory model to detect acquiescence and disacquiescence response styles (Park & Wu, 2019). Practical implications of IR-Tree models in examining response processes and how to link results with potential qualitative evidence are discussed.

Oral presentations session title:

Response Styles

Title

Extreme Responding under the IRTree and Multidimensional Nominal Response Models: Different Models, Different Outcomes

Author(s)

Martijn Schoenmakers, Jesper Tijmstra, Jeroen Vermunt, Maria Bolsinova

Tilburg University, Tilburg, Netherlands

Abstract

Extreme response style (ERS), the tendency of participants to select extreme item categories regardless of the item content, has frequently been found to decrease the validity of Likert-type questionnaire results (Moors, 2012). For this reason, various IRT models have been proposed to model ERS and correct for it. Comparisons of these models are however rare in the literature, especially in the context of cross-cultural comparisons, where ERS is even more relevant due to cultural differences between groups. To remedy this issue, the current paper examines two frequently used IRT models that can be estimated using standard software: a multidimensional nominal response model (MNRM) and a IRTree model. Studying conceptual differences between these models reveals that they differ substantially in their conceptualization of ERS. These differences result in different category probabilities between the models. To evaluate the impact of these differences in a multigroup context, a simulation study is conducted. Our results show that when the groups differ in their average ERS, the IRTree model and MNRM can drastically differ in their conclusions about the size and presence of differences in the substantive trait between these groups. An empirical example is given and implications for the future use of both models and the conceptualization of ERS are discussed.

Oral presentations session title:

Response Styles

1.8 Parallel sessions 11h30–13h00 Lecture room 1.2

Title

Machine-learning prediction of test item difficulty using item text wordings: Comparison of algorithms' and domain experts' predictive performance

Author(s)

Lubomír Štěpánek[1,2], Jana Dlouhá[1,3], Patrícia Martinková[1,4]

[1]Institute of Computer Science of the Czech Academy of Science, Prague, Czech Republic; [2]First Faculty of Medicine, Charles University, Prague, Czech Republic; [3]Faculty of Arts, Charles University, Prague, Czech Republic; [4]Faculty of Education, Charles University, Prague, Czech Republic

Abstract

Various properties of text wording of a given test item determine how difficult the item is for a test-taker. While the item difficulty is commonly estimated using item response theory (IRT) models based on test-takers' responses, information on item difficulty is encoded in its text and could be predicted using machine-learning algorithms.
In this work, we used text wordings of test items of the reading comprehension part of a test of English as a foreign language. For each item, we tokenized and lemmatized item text, removed stopwords, and calculated various features such as word counts, readability indices, lexical frequencies, and measures of item parts' similarity. Then, the resulting dataset containing text features in rows was enriched by item difficulty estimated using the Rasch model.
The item difficulty was predicted using multiple machine-learning supervised algorithms of regression task. Firstly, we applied regularization algorithms, i.e., LASSO, ridge regression, and elastic net, to select appropriate features, reduce dimensionality, and predict the (continuous) difficulty. Besides that, we employed support vector machines, regression trees and forests, and neural networks. Once we categorized the difficulty into disjunctive intervals, we switched the regression into a classification task, also applying the naïve Bayes classifier.
To compare the algorithms to each other and domain experts' difficulty predictions, we learned algorithms many times within cross-validation and estimated root mean square errors and predictive accuracies for each approach. Regularization algorithms in regression tasks and random forests in classification seemed to outperform other algorithms and predicted item difficulty similarly to domain experts.

Oral presentations session title:

Machine Learning

Title

Improving Questionnaire Efficiency: A Bayesian Networks, Jensen-Shannon, and Machine Learning Approach for Selecting Relevant Items and Assessing Symptomatology Risk.

Author(s)

Matteo Orsoni, Luca Tarasi, Sara Giovagnoli, Vincenzo Romei, Mariagrazia Benassi

University of Bologna, Bologna, Italy

Abstract

Scholars often develop a new questionnaire by merging two or more questionnaires previously validated. This procedure is often critical because of the selection of the items. Additionally, when the items are dichotomous only few procedures have been proposed in previous literature. We develop a novel method aimed to select the most relevant items in the case of dichotomous items questionnaires. By using the Autism Spectrum Quotient (AQ) (item N = 50) and the Schizotypal Personality Questionnaire (SPQ) (item N = 74) we develop a new questionnaire voted to recognize and distinguish subjects having Autistic and Schizotypal traits. 1081 adult healthy subjects participated in the study: 1.5% resulted having Autistic traits and 8.6% showed Schizotypal traits. The new method we propose for selecting the relevant items, consists of two steps: first the marginal probability and the JS distance distribution are evaluated then three machine learning (ML) models are compared - Decision Tree, Random Forest, and Support Vector Machine (SVM) – to achieve the best trade-off between the number of selected items and model performance. The SVM model performed best in selecting 25 variables, using an 80% variable exclusion threshold. Additionally, we developed an artificial neural network (ANN) model that predicted symptomatology risk with 85% test set accuracy and an 86% balanced F1 score.
To conclude, the Bayesian Networks and Jensen-Shannon divergence proved to be an effective method for selecting items in the case of dichotomous questionnaires.

Oral presentations session title:

Machine Learning

Title

Cross-validation methods for estimating the generalization error in psychological research

Author(s)

Diego Iglesias, Miguel A. Sorrel, Ricardo Olmos

Universidad Autónoma de Madrid, Madrid, Spain

Abstract

While traditionally most of psychological research has been focused on explanatory models, in recent years there has been an increasing interest in integrating machine learning predictive models. Regardless of whether the goal is to explain or predict psychological phenomena, any statistical model is built on a reduced set of observations drawn from the population of interest. Goodness of fit indices (i.e. in-sample error) tend to provide optimistic estimates of the true generalization error (i.e. out-of-sample error). Resampling methods such as cross-validation leverage all the information available in the sample to make a direct estimate of the generalization error. In the present study we studied the performance of three cross-validation methods hold out, k-folds and leave-one-out investigating their robustness against overfitting. The prediction of an independent variable from several predictors is adopted as the general setting. We conducted a simulation study in order to analyze the effect of sample size, model complexity, and signal-to-noise ratio on the accuracy of the generalization error estimates. Our results show that leave-one-out and k-folds provide the most accurate estimates. Whether the goal is to explain or to predict, since any statistical model is built on a reduced set of observations, accurate statistical techniques are required to prevent overfitting and obtain an unbiased estimation of the true generalization error. This study provides guidelines on how to proceed to apply procedures to better estimate the generalizability of the estimated model.

Oral presentations session title:

Machine Learning

Title

Detecting and describing interindividual heterogeneity using interpretable machine learning

Author(s)

Mirka Henninger, Rudolf Debelak, Yannick Rothacher, Carolin Strobl

University of Zurich, Zurich, Switzerland

Abstract

Many machine learning methods have the ability to automatically detect interaction effects and include them into the method's predictions. Therefore, they may become valuable tools to detect and describe interindividual heterogeneity in social science research. At the same time, many machine learning methods are so-called "black boxes" that do not allow us to see in what way the machine learning method has come to its prediction. In order to gain insights into the machine learning method, interpretation techniques have been proposed in recent years. They may support researchers to determine which predictor variables are involved in interaction effects or to describe in what way the predictors interact with each other.
In this talk, we present a selection of interpretation techniques that are specifically suited for detecting interactions between predictor variables, such as two-dimensional Partial Dependence and Accumulated Local Effect plots, Individual Conditional Expectation curves, and the Hamilton interaction statistic. Furthermore, we illustrate potential pitfalls and show via simulated examples how the interpretation techniques might miss interactions that are present or erroneously suggest interactions that are not present in the data. We believe that it is important to critically reflect on these interpretation techniques, as they will become powerful tools for describing interactions in social science research.

Oral presentations session title:

Machine Learning

1.9 Parallel sessions 11h30–13h00 Lecture room 1.3

Title

All Models Are Wrong, But ...: Testing Severely Theoretical Conjectures Regarding Risk Using Constrained Regression and Structural Equation Models

Author(s)

Keith Widaman

University of California, Riverside, CA, USA

Abstract

Developmental scientists have long used risk indices as shorthand indicators of barriers to optimal development. In the 1980s, Sameroff and colleagues identified 10 variables reflecting difficult environmental circumstances, dichotomized variables at notable points, and summed them to yield a 0 10 scale of caretaking risk or casualty. More recently, studies of gene-by-environment (G×E) interaction used single nucleotide polymorphisms (SNPs), coded 0, 1, 2 for the number of environmentally sensitive alleles, and summed to form a risk index across a set of SNPs. Thus, risk indices can be derived from environmental indices or genetic ones.
Typical regression approaches for analyzing risk indices embody exploratory methods, suffer from confirmation bias, test conjectures weakly, and lead to imprecise parameter estimates. The contrasting alternative is a falsificationist approach, testing conjectures severely to identify failures to corroborate predictions. This latter approach uses constrained modeling to obtain more precise answers to theoretical questions by testing them more directly. I will compare the typical regression approach (which uses unconstrained modeling) with the falsificationist approach that uses constrained modeling, providing contrasting analytic scripts using the lm, nls, and lavaan packages in R, PROC REG in SAS, among other programs.
Two empirical examples will be presented: a reanalysis of data from Sameroff et al., and an in-depth analysis of a G×E study. The typical regression and the falsificationist, constrained modeling approaches lead to different conclusions of non-trivial magnitude. Implications for research and theory and for the ways in which we analyze our data will be stressed.

Oral presentations session title:

Model Fit in Structural Equation Modeling (SEM)

Title

Some considerations about the model-size effect

Author(s)

Eric Klopp

Saarland University, Saarbrücken, Germany

Abstract

The model-size effect means that for models with latent variables, the model's asymptotically 2-distributed test statistic is upwardly biased depending on the size of the model, in particular, for small sample sizes. Drawing on statistical theory and the distinction between error of approximation and error of estimation, we reconsider some up-to-now neglected factors. We argue that the model-size effect appears only in correctly specified models. For correctly specified models, the expectancy of the test statistic equals the number of degrees of freedom. Because there is only error of estimation that decreases with increasing sample size, the test statistic is upwardly biased for small sample sizes. However, for misspecified models, there is both error of approximation and error of estimation. Therefore, the test statistic should be above the expectancy and, following statistical theory, should increase with increasing sample size and an increasing degree of misspecification. Thus, the model-size effect applies only to correctly specified models and means that for small samples size, the test statistic is above its expectancy. In this context, we also introduce a method to quantify a model's misspecification degree. Up-to-now neglected are the sizes of the manifest residual variances that also affect the test statistic: the smaller the manifest residual variance, the larger the upward bias in correctly specified models. However, for misspecified models, large manifest residual variances decrease the test statistic by masking the misfit, especially for small sample sizes. We demonstrate the above-mentioned behavior issues using simulations and discuss the implications.

Oral presentations session title:

Model Fit in Structural Equation Modeling (SEM)

Title

A Monte Carlo Examination of Two-Stage Approaches for Evaluating Structural Model Fit

Author(s)

Graham Rifenbark[1], Terrence Jorgensen[2]

[1]University of Connecticut, Storrs, USA; [2]University of Amsterdam, Amsterdam, Netherlands

Abstract

Fit indices enable evaluating the degree to which a structural equation model (SEM) can approximately reproduce observed (co)variances. To remove the influence of the measurement model when evaluating the structural component of an SEM, Hancock and Mueller (2011) proposed estimating a factor-covariance matrix to treat as observed data in a subsequent path model, from which standard fit indices can be calculated. The resulting structural fit indices (SFIs) possess inflated Type-I error rates when conventional thresholds for acceptable data-model fit are treated as critical values. We consider other two-step approaches for calculating SFIs that account for uncertainty about factor-covariance estimates: (a) analyzing plausible values of factor scores and (b) the Structural-After-Measurement (SAM) approach. We conducted a Monte Carlo simulation to explore how construct reliability, number of indicators-per-factor, measurement model (mis)specification, and sample size affected sampling distributions of SFIs, hypothesizing that plausible values and SAM could prevent measurement-model misspecification from biasing SFIs. Results show that under correct measurement-model specification, the main effect of construct reliability had a medium-to-large impact on popular fit indices using each approach, with better apparent fit with high construct reliability. No such effects were apparent under a misspecified measurement model, and average SFIs differed little across the three approaches. Differences between approaches became more apparent after inspecting the spread of SFI sampling distributions, which in turn affected Type I error rates when applying conventional rules of thumb as though they were critical values for test statistics.

Oral presentations session title:

Model Fit in Structural Equation Modeling (SEM)

Title

Tests of Model Fit for Structural Equation Models estimated from Finite Samples

Author(s)

Jonathan Helm

San Diego State University, San Diego, USA

Abstract

The common X2 test of model fit for a structural equation model relies on an asymptotic assumption, such that the sampling distribution of the test statistic (i.e., negative two multiplied by the difference between log-likelihoods for a nested and saturated model) approaches a X2-distribution (with degrees of freedom equal to the difference in the number of parameter estimates across the models) as sample size approaches infinity. Consequently, the test of model fit will be inaccurate for small sample sizes, and is known to produce an inflated Type-1 error rate (i.e., the correct model will be rejected more than the nominal [usually 5%; alpha = .05] rate). This presentation will identify and examine an alternative test statistic (and corresponding sampling distribution) which can better account for finite samples (i.e., whose derivation does not rely on asymptotics). A simulation will be presented which compares the newly defined statistic to other commonly implemented tests of fit, and an empirical example is presented.

Oral presentations session title:

Model Fit in Structural Equation Modeling (SEM)

1.10 Parallel sessions 16h00–17h30 Auditorium 1

Symposium Overview

Regularization in Structural Equation Models (SEM)

Author(s)

Sara van Erp[1], David Goretzko[1], Erik-Jan van Kesteren[1], Philipp Sterner[2]

[1]Utrecht University, Utrecht, Netherlands; [2]LMU Munich, Munich, Germany

Abstract

In any statistical model, we need to balance how well our model explains the phenomenon under investigation with the parsimony of this explanation. In practice, this often means choosing the right complexity for the model to balance the bias-variance tradeoff. In the context of regression models, regularization methods have been used successfully to balance the bias and variance, enabling automatic variable selection under some approaches. Regularization or penalization methods add a penalty term to the estimation procedure that serves as a failsafe to protect model parsimony and thereby avoid the problem that a model becomes too complex and as a result will not generalize to a new sample. In structural equation modeling, regularization methods have been applied for example to solve convergence problems and improve the performance of test statistics, to produce sparse loading matrixes in EFA, to automatically model cross-loadings and/or residual covariances in one step in CFA, to find relevant covariates in multiple indicators multiple causes (MIMIC) models, to identify potential mediators in mediation models, and to detect violations of measurement invariance. These possibilities of regularized SEM have even led Finch and Miller (2020) to argue that regularization methods might be used as default estimator in the context of SEM. In this symposium, we will discuss several recent developments in the context of regularized SEM as well as directions for future research.

Title

Tensorsem: An R Package for Structural Equation Modeling with Custom Penalties.

Author(s)

Erik-Jan van Kesteren

Utrecht University, Utrecht, Netherlands

Abstract

Regularization is a promising and flexible technique for imbuing structural equation models (SEM) with prior knowledge or inductive biases. For example, custom penalties can be used to allow some degree of residual correlation in factor analysis, or they can be used to select among many potential mediators in a path analysis. However, regularization and penalization move the parameter estimates away from the maximum likelihood estimates. Thus, the standard fitting methods used commonly for SEM are unavailable for these procedures.
The R package tensorsemprovides a flexible solution to this problem of estimating regularized SEM by borrowing two proven techniques from the deep learning field:
* computation graphs have been essential in obtaining automatic gradients for a wide variety of statistical learning methods; * adaptive first-order optimization methods have been shown to be capable of optimizing many non-standard likelihood-based loss functions.
The new and improved version 2 of the tensorsempackage implements both techniques natively in the R programming language, with a familiar user-interface (lavaan syntax). This allows researchers to add any type of custom penalty to any of the parameters in a structural equation model, and even to change the loss function altogether. The goal of the package is to accelerate the development of SEM models for modern, non-standard data problems.

Symposium title

Regularization in Structural Equation Models (SEM)

Title

Regularized Exploratory Factor Analysis – A Solution for Rotational Indeterminacy and Spurious Factors?

Author(s)

David Goretzko

Utrecht University, Utrecht, Netherlands

Abstract

Exploratory factor analysis (EFA) is a central statistical tool when developing measurement models for latent concepts and is widely used in psychological research. However, due to its exploratory nature, a user faces several important analysis decisions that affect the outcome. Among them, determining the number of factors and rotating an initial solution to obtain an interpretable factor structure (after the factor extraction) are certainly the most difficult decisions. Especially the problem of rotational indeterminacy for multivariate normal data may cause uncertainty in EFA users, as no data-driven way exists to decide among admissible rotation criteria. Recently, regularized EFA (REFA) has been proposed as an alternative to the common two-stage approach of estimating parameters and rotating the solution towards simple structure afterwards. Common maximum likelihood EFA can be extended by adding various penalties to the objective function to achieve sparse loading patterns. In this talk, we want to compare different penalization strategies and discuss potential advantages over common factor rotation but also possible shortcomings of REFA. Using simulated data, various penalties (e.g., LASSO, Ridge, ElasticNet, MC+) are evaluated regarding their ability to carve out the loading patterns of the data generating models and to select suitable indicators for latent factors. Exemplary data from different application contexts – psychological assessment, smartphone sensing and large panel data – will be used to illustrate the findings. In addition, REFA's potential to identify spurious factors that are the result of overfactoring will be debated.

Symposium title

Regularization in Structural Equation Models (SEM)

Title

Bayesian regularized structural equation modeling (SEM): Current capabilities and constraints

Author(s)

Sara van Erp

Utrecht University, Utrecht, Netherlands

Abstract

Classical regularized structural equation modeling (SEM) relies on optimization with a penalty function added to the usual estimation problem. An alternative to the classical approach is Bayesian regularized SEM in which the prior distribution serves as the penalty function. Specifically, Bayesian regularized SEM relies on shrinkage priors that have a high peak at zero with, ideally, heavy tails. The high peak at zero will pull small effects towards zero whereas the heavy tails allow large effects to escape this shrinkage behavior. Many different shrinkage priors exist, enabling great flexibility in terms of shrinkage behavior. In addition, advantages in terms of automatic uncertainty estimates, the possibility to include prior knowledge, and intuitive interpretation of the results have resulted in various applications of Bayesian regularization in SEM. However, the lack of user-friendly, general purpose software for Bayesian regularized SEM is lacking and as a result the full potential of Bayesian regularized SEM is not yet realized. In this presentation, I will review current applications of Bayesian regularized SEM and the corresponding software options. Through various illustrations, I will point out current capabilities as well as constraints with a discussion of the aspects future research and software development should focus on.

Symposium title

Regularization in Structural Equation Models (SEM)

Title

Exploratory Factor Analysis Trees and Regularization - Evaluating and Interpreting Measurement Invariance Between Multiple Covariates

Author(s)

Philipp Sterner[1], David Goretzko[2]

[1]LMU Munich, Munich, Germany; [2]Utrecht University, Utrecht, Netherlands

Abstract

Measurement invariance (MI) describes the equivalence of measurements of a construct across groups. To be able to meaningfully compare latent factor means between groups, it is crucial to establish MI. Although methods exist that test for MI, these methods do not perform well when many groups have to be compared or when there are no hypotheses about possible group constellations. In an exploratory factor analysis setting, an additional challenge is the choice of rotation method per group. The best method to use depends on the true population factor structure which is almost always unknown and likely different between groups. To address these issues, we first present a method called Exploratory Factor Analysis Trees (EFA trees) that are an extension to SEM trees (Brandmaier et al., 2013). EFA trees combine EFA with a model-based recursive partitioning algorithm that can uncover non-invariant subgroups in a data-driven manner. An EFA is estimated and then tested for parameter instability on multiple covariates (e.g., age, education, etc.) by a decision tree-based method. In this, EFA trees can simultaneously handle many categorical and continuous covariates. We then demonstrate how regularized EFA can be helpful to handle the rotational indeterminacy of factor solutions in the multigroup case. We compare the regularized solution to possible rotation choices and discuss future research ideas.

Symposium title

Regularization in Structural Equation Models (SEM)

1.11 Parallel sessions 16h00–17h30 Auditorium 2

Symposium Overview

Mixed-methods approaches to improve mental health research

Author(s)

Nekane Balluerka[1], Maria Dolores Hidalgo[2]

[1]University of the Basque Country UPV/EHU, San Sebastián, Spain; [2]University of Murcia, Murcia, Spain

Abstract

The World Health Organization defines mental health as "a state of mental well-being that enables people to cope with the stresses of life, realize their abilities, learn well and work well, and contribute to their community." Mental health is more than the absence of mental disorders. This symposium aims to provide mixed methodological approaches to improve conceptualization, assessment, and intervention in different contexts designed to enhance mental health research, services, and outcomes. Two of the contributions seek to improve the quality of data collection and assessment. The first compares results obtained online with those collected face-to-face using interviews and focus groups, to examine the strengths and weaknesses of each strategy and to propose the best methodological approach to respond to particular questions in optimal research contexts. The second proposes a procedure to identify and model careless responding, i.e., the lack of motivation to read or answer carefully questions of a questionnaire, and, in this way, improve the quality of assessment. In the third contribution, a mixed-methods approach is proposed to better understand patterns of health service delivery in a large governmental agency. In the fourth, a mixed methodological strategy is proposed with the aim of conceptualizing, rigorously evaluating, and implementing an intervention that follows the principles of recovery-oriented care of people with mental health problems. The symposium aims to underline the value of mixed-methods approaches for conducting high quality, evidence-based applied research focused on mental health.

Title

Exploring the equivalence of face-to-face and online data collection methods

Author(s)

Maite Barrios[1,2], Chuen Ann Chai[1], Juana Gómez-Benito[1], María Dolores Hidalgo[3]

[1]University of Barcelona, Barcelona, Spain; [2]Institute of Neuroscience, Barcelona, Spain; [3]University of Murcia, Murcia, Spain

Abstract

Online technologies have become more prevalent in research data collection since the Covid-19 pandemic. Online data collection methods have become increasingly popular due to their convenience, cost-effectiveness, and efficiency. However, the question remains: are online data collection methods as reliable and effective as traditional face-to-face data collection methods? The purpose of this study was to provide a comparative analysis of the characteristics of research face to face and online data collection methods with a view to assessing their usefulness as sources of data about the experience of a mental illness. We conducted face-to-face and online interviews and focus groups in different countries (e.g., Spain, India, Malaysia) as part of a mixed method study to assess functioning in people with schizophrenia. The Interviews and the focus groups were transcribed from recordings and meaningful units were coded by two independent coders following the International Classification of Functioning, Disability and Health (ICF) as a framework. Both face to face and online data collection methods were compared in terms of (a) the total number of words spoken by the interviewer and participant, (b) number of interviewer probes (e.g., number of active invitations for comment or clarifications of a statement or question), (c) number of meaningful units obtained from the interviewees responses, and (d) total duration of the interview or focus group. The findings of this study may have significant implications for researchers seeking to choose the most appropriate data collection method for their research project.

Symposium title

Mixed-methods approaches to improve mental health research

Title

Managing careless responding to improve data quality in Health and Social Sciences: A comparison of strategies

Author(s)

Inés Tomás, Ana Hernández, Vicente González-Romá

University of Valencia, Valencia, Spain

Abstract

Careless responding (CR) occurs when respondents fail to give sufficient attention to item content, resulting in poor quality data (Podsakoff et al., 2012). Thus, CR can impact the psychometric properties of the scales (PPS) and the substantive research results. The traditional recommendation to deal with CR is to eliminate the responses of careless respondents. More recently, other strategies have been proposed, such as introducing CR as a control variable or a moderating variable of the relationships of interest (e.g., Edwards, 2019). However, the adequacy of the different CR management strategies and their impact on the PPS and the results have not been empirically assessed yet. In this study we analyze the impact of the aforementioned CR management strategies (compared to doing nothing) on the PPS (reliability and validity) of an affective well-being scale (Kampf et al., 2020).
We analyzed a sample of 707 employees. 17.4% of them presented 1 or more errors in the 3 instructed items that were presented in a longer questionnaire. Reliability (omega) and factorial validity were obtained by fitting a series of CFA and MIMIC models by means of Mplus 8.8. Preliminary results indicate that eliminating careless respondents and introducing CR as a moderating variable of the relationship between items and factors (i.e., factor loadings) provide the best results. The last option seems to be a good way to keep the full sample considering the potential impact of CR. Research with mixed-methods will be helpful to understand the causes of CR and, thus, prevent it.

Symposium title

Mixed-methods approaches to improve mental health research

Title

A Mixed Methods Approach to Understanding Spending in Delivery of Health Improvement Services in a Large Governmental Agency

Author(s)

Keith Widaman, Jan Blacher

University of California, Riverside, California, USA

Abstract

The California Department of Developmental Services (DDS) provides services for persons with intellectual and developmental disabilities to over 110,000 persons annually. Initial analyses by DDS suggested that White recipients received substantially more services than members of other ethnic groups. To investigate this further, we outlined a three-pronged approach, analyzing (a) expenditure, (b) survey, and (c) focus group data.
Expenditures: We analyzed 5 years of expenditure data, which were highly skewed, with a large proportion of the sample receiving no expenditures in a given year. Thus, we created two variables: (a) a dichotomous variable that indicated whether an individual received any services or not, and (b) a continuous variable indexing the dollar amount of services received for those who received any services. When effects of key variables were controlled, the effect of ethnicity on dollar amount of services was nil – persons from all ethnic groups received about the same amount. However, logistic regression showed that members of other ethnic groups were less likely to receive services than were White persons.
Survey: Our survey of over 2,000 caregivers investigated whether differences across ethnic groups in odds of receiving services were due to bias by DDS or other factors (e.g., family/cultural matters).
Focus group: Our focus groups obtained more fine-grained information about caregiver interactions with DDS staff (e.g., satisfaction, conflicts).
We conclude with an overview of the ways in which data from mixed methods can be woven together to understand patterns of health service delivery in a large governmental agency.

Symposium title

Mixed-methods approaches to improve mental health research

Title

Revitalizing mental health recovery: A mixed methods approach for definition, assessment, and intervention

Author(s)

Georgina Guilera[1], Hernán Sampietro[2], Arantxa Gorostiaga[3], Nekane Balluerka[3]

[1]University of Barcelona, Barcelona, Spain; [2]ActivaMent Catalunya Associació, Barcelona, Spain; [3]University of the Basque Country UPV/EHU, Barcelona, Spain

Abstract

The recovery model in mental health shifts the focus from symptoms or functional adaptation to society towards the possibility of developing a satisfying life that aligns with the preferences and values of the individual. Our research team assumed the responsibility of promoting and implementing the recovery model in mental health policies in Spain, contributing to the conceptualization and evaluation of the recovery process, as well as the development of intervention programs and the study of their effectiveness. To tackle this significant challenge, we proposed a mixed methods approach. Firstly, we utilized the Delphi method to establish a consensus on the most relevant indicators that characterize the recovery process from three distinct perspectives: users, supporting network, and mental health professionals. The results allowed the development of a scale to assess the facilitators and barriers of the recovery process. Secondly, validation studies of various measurement instruments of recovery and related constructs (e.g., empowerment) were designed to offer the community a comprehensive set of tools that are applicable in the Spanish cultural context. Lastly, we designed an evaluation study of an intervention that aims to promote well-being and the development of a life project that is respectful of the preferences of people with mental health issues. In conclusion, we highlight and value the mixed methods methodology as an effective approach to address complex challenges in mental health, such as promoting the recovery paradigm in Spain.

Symposium title

Mixed-methods approaches to improve mental health research

1.12 Parallel sessions 16h00–17h30 Auditorium 3

Symposium Overview

Innovations in continuous-time statistical models for longitudinal change

Author(s)

Eduardo Estrada

Dpt. Social Psychology and Methodology. Univ. Autónoma de Madrid, Madrid, Spain

Abstract

Evaluating change over time is one of the most interesting problems in behavioral sciences. Most statistical models applied for that purpose traditionally define time in a discrete metric and quantify change from one given time point to the next. However, this approach involves several important problems. For example, the results are dependent on the chosen time interval. Furthermore, most psychological processes are assumed to exist also between observations, not only when they are measured.
Recently, continuous-time models have been proposed as a more general and powerful framework for characterizing change. Continuous-time models typically define the dynamics of change as a differential equation system. Very often, traditional discrete-time model can be considered a specific case of a more general continuous-time model, which is independent of the observed time-lagged, and more consistent with most theories in developmental, educational, and clinical psychology.
In this symposium, we present a set of cutting-edge advances in continuous-time dynamic modeling, and provide several perspectives on how they can be used to answer relevant substantive questions in psychology.

Title

Mixed-effects models with crossed random effects for individuals and variables using discrete and continuous time metrics

Author(s)

José Ángel Martínez-Huertas[1], Emilio Ferrer[2]

[1]National Distance Education University, Madrid, Spain; [2]University of California, Davis, Davis, USA

Abstract

Whilst most of the implementations of mixed-effects models have been done in univariate longitudinal processes, only occasionally they have been applied to bivariate or multivariate processes. In this presentation, we propose mixed-effects models with crossed random effects for individuals and variables as a novel and valuable alternative tool for the analysis of multivariate longitudinal data. These models can consider different sources of variability simultaneously and can easily accommodate discrete and continuous time metrics. In our proposal, we consider crossed random effects for individuals and variables, which are fully crossed: all the variables are measured in all the individuals, except for cases of missing data. Here, we illustrate the use of these models in two types of longitudinal studies based on balanced and unbalanced data: panel studies and cohort-sequential designs, respectively. We conclude that mixed-effects models with crossed random effects provides relevant information about the developmental trajectories of individuals and variables in multivariate longitudinal data under either type of data condition. Some of these results were recently published (https://doi.org/10.1080/10705511.2022.2108430).

Symposium title

Innovations in continuous-time statistical models for longitudinal change

Title

Ensembles of continuous-time SEM trees using structural change tests

Author(s)

Pablo F. Cáncer[1], Manuel Arnold[2], Eduardo Estrada[1], Manuel Voelkle[2]

[1]Dept. Social Psychology and Methodology. Univ. Autónoma de Madrid, Madrid, Spain; [2]Dept. Psychology. Humboldt-Universität zu Berlin, Berlin, Germany

Abstract

Purpose. Model-based recursive partitioning has been gaining traction in psychological research. The technique finds similar individuals in heterogeneous data sets and identifies the most important predictors of group differences in the process. In the past decade, structural equation models (SEM) have been almost entirely partitioned using the semtree software package, leading to so-called SEM trees and forests. Recently, score-based covariate testing has been implemented into semtree, drastically improving runtime and making the partitioning of more complex models possible. In the present work, we extended this approach to continuous-time models. Unlike discrete-time (DT) models, CT models adapt effortlessly to longitudinal data observed with different time intervals between measurements. Thus, our resulting approach, which we call score-based CT-SEM trees and forests, is well suited to deal with heterogeneity between individuals and measurement occasions. However, it is uncertain whether CT-SEM forests will be feasible in terms of computation time and whether their performance will be acceptable for the empirical practice.
Method. To answer these questions, we conducted a Monte Carlo study to evaluate the performance of CT-SEM forests under a broad set of empirically relevant conditions. We also illustrated the application of a CT-SEM forest using empirical data from the Survey of Health, Ageing, and Retirement in Europe (SHARE).
Results and discussion. We discuss the most relevant findings, elaborate on the strengths and limitations of the proposed algorithm, and comment on current challenges and future lines of research in the context of between-individual differences in change.

Symposium title

Innovations in continuous-time statistical models for longitudinal change

Title

Recovering trajectories of bivariate dynamics in accelerated longitudinal designs from a continuous time approach

Author(s)

Nuria Real-Brioso, Pablo F. Cáncer, Eduardo Estrada

Dept. Social Psychology and Methodology. Univ. Autónoma de Madrid, Madrid, Spain

Abstract

Accelerated longitudinal designs (ALDs) provide an opportunity to capture long developmental periods with a smaller number of assessments in a shorter time framework. Prior literature has investigated discrete- and continuous-time approaches and their ability to recover univariate developmental processes from ALD data. However, some processes, such as cognitive and cortical development, are intercorrelated as they unfold over time, necessitating the use of bivariate models to be analyzed. To date, such models have not been studied in the context of ALDs. We conducted a Monte Carlo simulation study to evaluate the performance of continuous-time bivariate latent change score models in recovering intercorrelated trajectories under different ALD sampling conditions. We provide some insights and guidance into the application of bivariate models in ALDs for the study of complex developmental processes.

Symposium title

Innovations in continuous-time statistical models for longitudinal change

Title

Examination of the Damped Linear Oscillator model for the idiographic study of affect dynamics in clinical psychology

Author(s)

Mar J.F. Ollero[1], Pablo F. Cáncer[1], Michael D. Hunter[2], Eduardo Estrada[1]

[1]Dept. Social Psychology and Methodology. Univ. Autónoma de Madrid, Madrid, Spain; [2]Dept. Human Development and Family Studies, Pennsylvania State Univ., State College, PA, USA

Abstract

People show stable differences in the way their affect fluctuates over time. Within the general framework of dynamical systems, the damped linear oscillator (DLO) model has been proposed as a useful approach to study affect dynamics. The DLO model can be applied to repeated measures provided by a single individual, and the resulting parameters can capture relevant features of the person's affect dynamics. Focusing on negative affect, we provide an accessible interpretation of the DLO model parameters in terms of emotional lability, resilience, and vulnerability. We conducted a Monte Carlo study to test the DLO model performance under different empirically relevant conditions in terms of individual characteristics and sampling scheme. We used State-Space Models (SSM) in continuous-time. The results show that, under certain conditions, the DLO model is able to recover the parameters underlying the affective dynamics of a single individual accurately and efficiently. We discuss the results and the theoretical and practical implications of using this model, illustrate how to use it for studying psychological phenomena at the individual level, and provide specific recommendations on how to do collect data for this purpose.

Symposium title

Innovations in continuous-time statistical models for longitudinal change

1.13 Parallel sessions 16h00–17h30 Auditorium 4

Title

Bayesian evidence synthesis for informative hypotheses: An aggregation tool for evidence from conceptual replications

Author(s)

Irene Klugkist, Thom Volker

Utrecht University, Utrecht, Netherlands

Abstract

Bayesian Evidence Synthesis (BES) is a method that aggregates levels of evidence from multiple studies. BES is highly flexible because it aggregates at the level of hypotheses instead of on the level of data or model parameters (as, for instance, in meta-analysis). It can therefore aggregate results of highly diverse studies (e.g., with different designs, statistical models, and/or variables) as, for instance, obtained from conceptual replication.
In this study we applied BES to the evaluation of informative hypotheses. The evaluation of informative hypotheses in a single study, using Bayes factors, has been well established in the literature. Aggregating evidence from such evaluations using BES is, however, still rather new and therefore requires methodological investigation of its performance and potential limitations. This study will present results and conclusions from several simulation studies.

Oral presentations session title:

Bayesian Analysis

Title

A Similarity-Weighted Informative Prior Distribution for Bayesian Multiple Regression Models

Author(s)

Christoph Koenig

Goethe University, Frankfurt, Germany

Abstract

Specifying accurate informative prior distributions is a question of carefully selecting studies that comprise the body of comparable background knowledge. Results of previous studies, however, are heterogeneous, and not all available results should contribute equally to an informative prior distribution. Current approaches to account for heterogeneity by weighting informative prior distributions, such as the power prior and the meta-analytic predictive prior are either not easily accessible or incomplete. To complicate matters further, in the context of Bayesian multiple regression models there are no methods available for quantifying the similarity of a given body of background knowledge to the focal study at hand. Consequently, the purpose of this study is threefold. We first present a novel method to combine the aforementioned sources of heterogeneity in the similarity measure . This method is based on a combination of a propensity-score approach to assess the similarity of samples with random- and mixed-effects meta-analytic models to quantify the heterogeneity in outcomes and study characteristics. Second, we show how to use the similarity measure as a weight for informative prior distributions for the substantial parameters (regression coefficients) in Bayesian multiple regression models. Third, we investigate the performance and the behavior of the similarity-weighted informative prior distribution in a comprehensive simulation study, where it is compared to the normalized power prior and the meta-analytic predictive prior. The similarity measure and the similarity-weighted informative prior distribution as the primary results of this study provide applied researchers with means to specify accurate informative prior distributions.

Oral presentations session title:

Bayesian Analysis

Title

Bayesian latent growth curve modeling with criminological panel data

Author(s)

Jasper Bendler, Jost Reinecke

Bielefeld University, Bielefeld, Germany

Abstract

The use of latent growth models is particularly useful for analyzing development trajectories, and this method is widely used in criminological research. The recent emergence of Bayesian structural equation models and the associated emergence of Bayesian growth models offer many new possibilities for this kind of research. In the presentation, examples of Bayesian latent growth models based on criminological panel data will be given. Based on this, the possibility of Bayesian latent class growth analyses (LCGA) against the background of heterogeneous growth trajectories will also be presented. The basis for this is the long-term panel study crime in the modern city (CrimoC) with a preliminary survey in Muenster (Germany) including 4 waves and a subsequent main survey in Duisburg (Germany) including 13 waves and over 3,000 participants. The substantive focus of the presented models will be the emergence and decline of youth criminality during adolescence and young adulthood. Advantages and disadvantages of the Bayesian approach compared to the classical frequentist approach will also be discussed.
References: Kessler, G. (2020): Delinquency in Emerging Adulthood: Insights into Trajectories of Young Adults in a German Sample and Implications for Measuring Continuity of Offending. In: Journal of Developmental and Life-Course Criminology, 6(4), 424-447.
Erdmann, A. Reinecke, J. (2019): What Influences the Victimization of High-Level Offenders? A Dual Trajectory Analysis of the Victim-Offender Overlap From the Perspective of Routine Activities With Peer Groups. In: Journal of Interpersonal Violence, 36(17-18), NP9317–NP9343.

Oral presentations session title:

Bayesian Analysis

Title

Bayesian Evaluation of N=1 Studies

Author(s)

Herbert Hoijtink

Utrecht University, Utrecht, Netherlands

Abstract

The Bayes factor and corresponding posterior model probabilities can be used to evaluate a set of competing hypotheses. The set can contain a null hypothesis, one or more informative hypotheses (e.g., b1 > 0, b2 >0 & b3 >0, where the b's denote regression coefficients), and the complement of the union of these hypotheses. First of all, Bayes factor and posterior model probabilities will introduced. Subsequently, two N=1 studies will be introduced. The first concerns tracking three outcome variables for a patient that is receiving trauma therapy for a period of thirteen weeks. The hypotheses of interest are H0: b1 = 0, b2 = 0 & b3 = 0, that is there is no effect of the therapy on the outcome measures; H1: b1 < 0, b2 < 0 & b3 > 0, that is, two outcome measures decrease and one increases as a result of the therapy; and, the complement of the union of both hypotheses, that is, something else is going on. It will be shown that thirteen measurements of three outcome variables is enough to evaluate these hypotheses. The second example concerns tracking a family (parents and two children) during the 52 weeks in which they receive counseling. In each week the inter-family experience level of violence is recorded for each family member. Hypotheses regarding the development of "experience level of violence" will be formulated and evaluated. The presentation is concluded with a short discussion.

Oral presentations session title:

Bayesian Analysis

1.14 Parallel sessions 16h00–17h30 Lecture room 1.2

Title

Addressing missing values with a Substantive-Model-Compatible approach in crossed random-effects models

Author(s)

Susana Sanz[1], Ricardo Olmos[1], Carmen García[1], José Ángel Martínez-Huertas[2]

[1]Universidad Autónoma de Madrid, Madrid, Spain; [2]Universidad Nacional de Educación a Distancia, Madrid, Spain

Abstract

In recent years, there has been a significant focus on addressing the issue of missing data in multilevel models (MLM). A fully Bayesian approach using Substantive-Model-Compatible (SMC) based methods has emerged as a promising strategy for dealing with missing data in MLM, as modern methods do not perform well in certain scenarios (e.g., when non-linear effects, such as interaction levels between two level-1 variables, or cross-level interactions are present). However, there has been little research conducted on MLM when two random factors are crossed rather than nested. Several empirical investigations, including those in the domains of psycholinguistics and education, utilize multilevel models that incorporate crossed-random factors. The aim of this research is to demonstrate how to handle missing data in Crossed Random-Effects Models using an SMC approach, as opposed to using listwise deletion. A simulation study was conducted to evaluate this approach and the risks associated with not considering all the random structures present in the data, by treating both subjects and tasks as random effects, as opposed to only considering the subjects-related effects. The findings show that the SMC approach results in non-biased estimations, whereas conventional missing data methods tend to underestimate fixed effects. Furthermore, it is essential to consider the correct random structure when dealing with missing values in this type of multilevel model.

Oral presentations session title:

Missing Values

Title

Tackling challenges in data synthesis: missing data handling in latent variable models with continuous and categorical indicators.

Author(s)

Lihan Chen, Carl Falk, Milica Miocevic

McGill University, Montreal, Canada

Abstract

Data synthesis is a flexible tool in empirical research, but pooling multiple datasets generally results in missingness and a mix of continuous and categorical items. Since this type of missingness is determined by data sources that vary in key population characteristics, the missing at random assumption is typically tenable, and modern missing data techniques such as full information maximum likelihood (FIML) can be applied. However, FIML assumes continuous variables only. Alternatively, since categorical data can be analyzed using least square approaches based on polychoric correlations, a three-stage approach can be used, applying deletion methods during the estimation of correlations; but this method may only perform well when data are missing completely at random. This leads to two commonly accessible approaches for data synthesis, 1) treat discrete data as continuous to perform FIML, or 2) use three-stage missing data handling for categorical data. While both approaches have theoretical drawbacks, their performances have not been systematically evaluated in this context. We performed simulation studies on these missing data approaches under mixed continuous and ordinal data. The simulation included a confirmatory two-factor model and a mediation model with latent variables, under different number of indicators, missing data mechanisms, rate of missing data, proportion of ordinal variables, various nonnormality conditions, etc. Preliminary results indicate treating data as continuous using FIML outperforms the categorical approach in most scenarios we investigated; we explored cases where the continuous approach breaks down. We make recommendations to empirical researchers performing data synthesis based on these results.

Oral presentations session title:

Missing Values

Title

Towards a standardized evaluation of imputation methodology: potential pitfalls in simulation studies and a proposed course of action

Author(s)

Hanne Oberman, Gerko Vink

Utrecht University, Utrecht, Netherlands

Abstract

Developing new imputation methodology for incomplete data has become a very active field. Unfortunately, there is no consensus on how to perform simulation studies to evaluate the properties of imputation methods. In part, this may be due to different aims between fields and studies. For example, when evaluating imputation techniques aimed at prediction, different objectives may be formulated than when statistical inference is of interest. The lack of consensus may also stem from different personal preferences or scientific backgrounds. All in all, the lack of common ground in evaluating imputation methodology may lead to suboptimal use of missing data methods in practice. We propose a move towards a standardized evaluation of imputation methodology. To demonstrate the need for standardization, we highlight a set of possible pitfalls that bring forth a chain of potential problems in the objective assessment of the performance of imputation routines. Additionally, we suggest a course of action for simulating and evaluating missing data problems. Our suggested course of action is by no means meant to serve as a complete cookbook, but rather meant to incite critical thinking and a move towards objective and fair evaluation of imputation methodology.

Oral presentations session title:

Missing Values

Title

Performance of stacked multiple imputations in different model selection approaches for cross-sectional networks

Author(s)

Kai Jannik Nehler, Martin Schultze

Goethe Universität, Frankfurt, Germany

Abstract

Psychological networks are used to illustrate relations between aspects of psychological constructs or disorders. Observed variables are represented as nodes, which are connected by edges indicating the strength and sign of their relationship by partial correlations. Currently, a broad collection of research focuses on model section approaches to identify true edges. However, the impact of missing values, a common problem in psychological research, is rarely investigated for cross-sectional network analysis. While EM-algorithms and traditional deletion techniques are implemented in most common network-analysis software, multiple imputation is not yet readily available. This may be due to the problem of combining model selection approaches with multiple data sets generated by this technique. Here we present a possible solution that is adapted from regularized multiple regression: performing analysis on the stacked dataset. In this talk, we present the results of a simulation study investigating the performance of stacked multiple imputation in combination with two different model selection techniques: traditional glasso regularization (Friedman et al., 2007) and Williams' (2020) nonconvex regularization, which mimics best subset selection. We varied network size, number of observations, percentage of missingness, and the missing data mechanism to determine the feasibility of this missing data handling approach in a variety of settings. Performance criteria include sensitivity and specificity in detecting truly non-zero edges, bias in strength centrality, and loss in the precision matrix.

Oral presentations session title:

Missing Values

1.15 Parallel sessions 16h00–17h30 Lecture room 1.3

Title

Parametric and nonparametric propensity score estimation in multilevel observational studies

Author(s)

Marie Salditt, Steffen Nestler

University of Münster, Münster, Germany

Abstract

There has been growing interest in using nonparametric machine learning approaches for propensity score estimation in order to foster robustness against misspecification of the propensity score model. However, the vast majority of studies focused on single-level data settings, and research on nonparametric propensity score estimation in clustered data settings is scarce. In this article, we extend existing research by describing a general algorithm for incorporating random effects into a machine learning model, which we implemented for generalized boosted modeling (GBM). In a simulation study, we investigated the performance of logistic regression, GBM, and Bayesian additive regression trees (BART) for inverse probability of treatment weighting (IPW) when the data are clustered, the treatment exposure mechanism is nonlinear, and unmeasured cluster-level confounding is present. For each approach, we compared fixed and random effects propensity score models to single-level models and evaluated their use in both marginal and clustered IPW. We additionally investigated the performance of the standard Super Learner and the balance Super Learner. The results showed that when there was no unmeasured confounding, logistic regression resulted in moderate bias in both marginal and clustered IPW, whereas the nonparametric approaches were unbiased. In presence of cluster-level confounding, fixed and random effects models greatly reduced bias compared to single-level models in marginal IPW, with fixed effects GBM and fixed effects logistic regression performing best. Finally, clustered IPW was overall preferable to marginal IPW and the balance Super Learner outperformed the standard Super Learner, though neither worked as well as their best candidate model.

Oral presentations session title:

Causal Inference

Title

The Parametric g-Formula for Latent Markov Models

Author(s)

Felix J. Clouth[1], Maarten J. Bijlsma[2], Steffen Pauws[1], Jeroen K. Vermunt[1]

[1]Tilburg University, Tilburg, Netherlands; [2]Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, Netherlands

Abstract

Post-treatment confounding poses a major challenge for causal inference on longitudinal data. When data is collected within an observational study design, there will be a self-selection process into the treatment groups and the causal effect of a treatment on an outcome of interest will be confounded. This problem intensifies with repeatedly measured outcomes when individuals are allowed to switch between treatment groups. E.g., a treatment might not have the expected effect on the outcome for some individuals and treatment might therefore be adjusted at follow-up. It is also possible that treatment affects some time-varying confounders which then affect treatment allocation at follow-up. If this happens, the average treatment effect will be confounded. To solve this problem of post-treatment confounding, the parametric g-formula has been proposed. In the social sciences, however, this framework needs to be extended as we are often confronted with outcomes that are not directly observable but are measured through indicators, i.e., are latent. For longitudinal data, latent Markov models are a common choice for modeling such outcomes. In this talk, I will present an extension of the parametric g-formula for unobserved outcomes. In a stepwise approach, we first estimate the measurement part of the latent Markov model. With the measurement model fixed, we then combine the estimation of the Markov chain with the parametric g-formula to account for time-varying confounding.

Oral presentations session title:

Causal Inference

Title

The rocky road from a randomized experiment to causal inference:The effect of treating childhood anxiety on young adult substance use disorders.

Author(s)

Lissette M. Saavedra[1], Antonio A. Morgan-López[1], Stephen G. West[2,3], Margarita Algeria[4], Wendy K. Silverman[5]

[1]Research Triangle Institute, Research Triangle Park, NC, USA; [2]Arizona State University, Tempe, AZ, USA; [3]Freie Universität jBerlin, Berlin, Germany; [4]Massachusetts General Hospital, Boston, MA, USA; [5]Yale University, New Haven, CT, USA

Abstract

The randomized trial is viewed as the "gold standard" of research designs for evaluating interventions. Yet, in research evaluating clinical and medical interventions, control patients are often waitlisted and later crossed over to treatment. To support the strongest possible causal inference about the long-term effect of Cognitive Behavioral Therapy (CBT) in children experiencing anxiety disorders prior to age 10 on reducing substance use disorders in young adults, several methodological challenges needed to be addressed. (1) An appropriate comparison group (an epidemiological sample with untreated participants) had to be located. (2) Participants in the intervention study needed to be equated using propensity score matching with participants in the epidemiological study. (3) Outcome measures in the intervention and epidemiological studies overlapped, but were not identical. These overlapping measures needed to be harmonized using integrative data analysis procedures, here moderated non-linear factor analysis. (4) Recognizing that clinical disorders are not normally distributed, the latent variables produced by the harmonization procedures could not assume underlying normal distributions. The latent outcome variable was produced by an adaptation of a 2-class mixture model developed by Wall et al. (2015). (5) Some participants for whom CBT was unsuccessful and participants in the epidemiological sample may have received other post-intervention treatments outside of the study. This potential bias was addressed through a separate analysis that removed the effect of the alternative treatment. Addressing these challenges reduces potential bias in accounting for the association between treatment and outcome supporting improved causal inference.

Oral presentations session title:

Causal Inference

Title

Identifying Causal Effects for Key Parameters in Latent State-Trait Models

Author(s)

Fabian Münch[1], Christian Gische[2], Manuel C. Voelkle[2], Tobias Koch[1]

[1]Friedrich-Schiller-Universität, Jena, Germany; [2]Humboldt-Universität zu Berlin, Berlin, Germany

Abstract

Latent state-trait (LST) models are widely used for analyzing complex longitudinal data that may exhibit a stable or changing trait, occasion-specific deviations from the trait, as well as autoregressive (carry-over) effects across measurement occasions. LST models have also been used as a basis for the identification of average, conditional, and individual causal effects in observational studies. However, traditional LST models are limited in their capacity for identifying causal effects in terms of key model parameters that characterize additive (level) trait change, multiplicative trait change, or trait change due to past experiences; furthermore, it is not fully clear how to account for time-stable and time-varying covariates in the analysis. In this talk, we present a Bayesian moderated nonlinear latent state-trait (MNLST) approach (Oeltjen et al., 2023) and show how it can be useful in identifying causal effects of interventions on additive and multiplicative trait change parameters in LST models, building upon the stochastic theory of causal effects by Steyer and colleagues. We illustrate the MNLST framework using data of major German panel studies (SOEP and Pairfam) and discuss model evaluation and hypothesis testing regarding model parameters in a Bayesian setting.

Oral presentations session title:

Causal Inference

1.16 Poster session 1 14h00–15h00

Title

Optimal Sample Size for the Variance Ratio: Considering Hypothesis Testing, Confidence Intervals, and Prediction Intervals

Author(s)

Weiming Luh[1], JiinHuarng Guo[2]

[1]National Cheng Kung University, Tainan City, Taiwan; [2]National Pingtung University, Pingtung, Taiwan

Abstract

Comparing population variance ratios has many applications and is routinely performed. Such comparison constitutes a classic problem and so is of interest to researchers. However, the statistical literature on sample size planning concerning both statistical power and precision is inadequate. It may acquire enough sample sizes to reject the null hypothesis (event rejection), to encompass the true parameter for a $100(1-\alpha)\%$ two-sided CI (event validity), and/or the two-sided CI width can achieve a desired width (event width). In addition, the use of unequal sample size allocation offers a number of advantages for cost-effectiveness but this issue has received limited attention. Thus, this article develops the optimal sample size to unify the hypothesis test of the variance ratio and/or the construction of a confidence interval. In addition to that, the sample size needed is also considered for a prediction interval. To fill the research gap, nine probabilities of combined events of rejection, validity, and/or width were specified, and an exhaustive search is used to develop several R Shiny apps for easy application. An example of a blood hypertension study is also illustrated. The present study provides a complete mathematical framework that was not treated in much detail in the past. Moreover, the framework leads directly to the proposed easy-to-use apps, and statistical power, precision, and sampling cost can be accurately implemented by statistical practitioners in many disciplines.

Title

Factor structure of the Multidimensional Body-Self Relations Questionnaire (MBSRQ) in a general Spanish sample: Concordance between four-factor and two-factor models.

Author(s)

Sergio Navas-león[1], Milagrosa Sánchez-Martín[1], Luis Morales Márquez[1], Ana Tajadura-Jiménez[2]

[1]Universidad Loyola Andalucía, Sevilla, Spain; [2]Universidad Carlos III de Madrid, Madrid, Spain

Abstract

The factor structure of the MBSRQ is under debate due to the limitations of previous research methods such as principal component analysis (PCA). Although the four-dimensional 38-item version of the MBSRQ is widely used in Spain, recent studies suggest a more efficient two-factor 15-version item. However, no studies have directly compared both versions with a Spanish population sample. This study aims to compare the classic four-factor model and the two-factor model to contribute to a better understanding of its validity and inform future research and practice in the assessment of body image in Spanish-speaking populations. A sample of 1012 participants was selected using stratified random sampling. The factor structure was investigated using CFA through WLSMV. ME/I was evaluated through the change in CFI/ RMSEA. The four-factor model showed poor psychometric properties [CFI=.696; TLI=.676; SRMR=.098; RMSEA=.109 (90% CI=.107-.111)]. Conversely, the second-factor model version showed an adequate fit [CFI=.925; TLI=.917; SRMR=.067; RMSEA=.050 (90% CI=.048-.053)]. For this version, values for reliability, as well as convergent and discriminant validity, were satisfactory. ME/I models were upheld, confirming the invariance across sex. CFA results did not support the original factor structure. A second-factor model fits the data reasonably well with factors corresponding to subjective importance of physical appearance and subjective importance of physical fitness. Remarkably, the two-factor model was equivalent for men and women, being a useful tool for practitioners to design interventions. Future studies should examine the predictive/concurrent validity.

Title

Development and pre-test of the Juvenile Victimization Questionnaire (JVQ) self-report version for children between 8 and 12 years old.

Author(s)

Ana Greco[1,2,3], Irene Montiel[1,2,3], Noemí Pereda[4,5,3]

[1]Universitat Oberta de Catalunya, Barcelona, Spain; [2]VICRIM - Grup de Recerca en Justícia Penal, Barcelona, Spain; [3]Grup de Recerca en Victimoloia Aplicada i Empírica, Barcelona, Spain; [4]Universitat de Barcelona, Barcelona, Spain; [5]Grup de Recerca en Victimització Infantil i Adolescent, Barcelona, Spain

Abstract

Research on child victimization has traditionally relied on proxy-informants or retrospective questionnaires. Nowadays, evidence has shown that asking directly to children about their lives has unique value, especially in sensitive issues like violent experiences (Devries et al., 2015). However, instruments to ask children if they have ever experienced violence are designed as interviews, do not report adequate psychometric properties or are addressed to children over 11 years old (y.o., Mathews et al., 2020; Meinck et al., 2022). Our aim was to adapt the most sound, comprehensive, and used instrument worldwide to measure violence against children, i.e., the Juvenile Victimization Questionnaire (JVQ, Finkelhor et al., 2005) to a self-report version suitable in Spanish for children between 8 and 12 y.o. To do so, methodological recommendations about the target population (e.g., limited responses options, including illustrations) were considered to create a 15-item self-report version of the JVQ including five victimization modules (i.e., caregivers, peer, sexual, electronic victimization and exposure to violence). Content validity (i.e., relevance, comprehensiveness, comprehensibility) and responsiveness was assessed through an experts review (n = 38) and tested through cognitive interviews and focus groups with children (n = 25). Both instances also tested understanding, adequateness and feasibility. Criterion-related validity evidence was gathered through a pilot test (n > 300) in which victimization experiences were correlated with psychological wellbeing. We report the psychometric properties of this JVQ version, which makes it possible to include the voices of children between 8 and 12 y.o. in child victimization research.

Title

Violation of Sphericity and normality in repeated measures designs

Author(s)

María J. Blanca[1], Jaume Arnau[2], F. Javier García-Castro[3], Rafael Alarcón[1], Roser Bono[2]

[1]University of Malaga, Malaga, Spain; [2]University of Barcelona, Barcelona, Spain; [3]Universidad Loyola Andalucía, Sevilla, Spain

Abstract

The assumptions of normality and sphericity must be fulfilled for repeated measures analysis of variance. A number of statistical procedures have been proposed for those cases where these assumptions are violated, the most common being the Greenhouse-Geisser (F-GG) and Huynh-Feldt (F-HF) adjustments, both of which are intuitive, easy to use and available in most statistical software. Although there is a great deal of research on the robustness of these procedures, the results are somewhat inconsistent and there are no clear guidelines for applied researchers, hence the need for further studies. The aim of this study was to analyse the performance of the F-statistic, F-GG and F-HF in terms of Type I error, with designs including 3 repeated measures, non-normal data, values ranging from the lower to its upper limit and sample sizes from 10 to 300. Non-normal data are represented by slight, moderate and severe deviation from normality, including both unknown and known distributions such as the lognormal distribution (= 1, = 0.5). The results showed that both F-HF and F-GG are robust alternatives to F when sphericity and normality are violated with values below .90, this being the case for all conditions studied except with samples as small as 10, extreme violation of sphericity (.50) and severe violation of normality (lognormal distribution). Further studies are needed to analyse robustness when sphericity is violated, with other non-normal distributions and with a greater number of repeated measures. This research was supported by grant PID2020-113191GB-I00 from the MCIN/AEI/10.13039/501100011033.

Title

One-way ANOVA effect size estimators under non-normality

Author(s)

F. Javier García-Castro[1], María J. Blanca[2], Jaume Arnau[3], Rafael Alarcón[2], Roser Bono[3]

[1]Universidad Loyola Andalucía, Sevilla, Spain; [2]University of Malaga, Malaga, Spain; [3]Univesity of Barcelona, Barcelona, Spain

Abstract

One-way ANOVA uses the F-statistic to determine whether group means of the dependent variable are equal. This test should be complemented with measures of effect size, defined as the degree to which the phenomenon occurs in the population. Although simulation studies have been used to analyse the bias, precision and accuracy of the three main effect size estimators (eta-squared, epsilon-squared, and omega-squared), most such studies have focused on bias with a normal distribution and small sample size. The aim of the present study was to investigate the impact of non-normality on the bias, precision and accuracy of these estimators. We considered a four-group, between-subject design with group size from 10 to 100, a linear pattern of means, and small (f = .10) and large (f = .40) magnitude effect size. The normal distribution and distributions with slight, moderate and severe deviation from normality were also included. Results showed that eta-squared was the most biased and the least accurate estimator in all conditions, especially with smaller samples and severe deviation from the normal distribution. All estimators yielded similar precision, which decreased with severe deviation from normality. Accuracy of all estimators was better for larger samples, with eta-squared being the least accurate estimator and the most affected by non-normality. Overall, the results for non-normal distributions were similar to those for the normal distribution, although there was an increase in bias and a decrease in precision and accuracy with severe deviation from normality. This research was supported by grant PID2020-113191GB-I00 from the MCIN/AEI/10.13039/501100011033.

Title

Power of generalized linear mixed models for mixed designs with binary data: A simulation study

Author(s)

Roser Bono[1], Rafael Alarcón[2], Jaume Arnau[1], F. Javier García-Castro[3], María J. Blanca[2]

[1]University of Barcelona, Barcelona, Spain; [2]University of Malaga, Malaga, Spain; [3]Universidad Loyola Andalucía, Sevilla, Spain

Abstract

This study examined the statistical power of generalized linear mixed models (GLMM). These models estimate fixed and random effects, and they are especially useful when the dependent variable is binary and when it involves repeated measures. Monte Carlo simulation was used to analyse the power of GLMM (value equal to or greater than .80) in mixed designs with two levels both between and within factors. The variables manipulated in the simulation studies were as follows: sample size (N from 24 to 492), coefficient of group size variation (Δn = 0.16, 0.33 and 0.50) and effect size (small and medium). The results for the time and interaction effects indicated that power of .80 or greater was achieved: (a) for balanced groups and small effect size with N = 84; (b) for balanced groups and medium effect size with N = 36; (c) for unbalanced groups and small effect size with N = 84 and Δn = 0.16, and with N = 108 for any value of Δn; and (d) for unbalanced groups and medium effect size with N = 36 and Δn = 0.16 and 0.33, and with N = 48 for any value of Δn. In conclusion, with medium effect and sample sizes, the amount of inequality in group sample sizes does not affect the power of GLMM. This research was supported by grant PID2020-113191GB-I00 from the MCIN/AEI/10.13039/501100011033.

Title

Development and psychometric validation of a model describing users' willingness to delegate to digital assistants

Author(s)

Ekaterina Svikhnushina[1], Marcel Schellenberg[2], Anna Niedbala[2], Iva Barisic[2], Jeremy Miles[3]

[1]EPFL, Lausanne, Switzerland; [2]Google, Zurich, Switzerland; [3]Google, Los Angeles, USA

Abstract

As digital assistants (DAs), such as Apple's Siri, Amazon's Alexa or Google Assistant, continue to evolve and gain new users, it is important to understand and prioritize the factors that drive consumer adoption of this technology. This study aimed to develop a model of users' behavioural intentions to delegate tasks to their DA and evaluate its psychometric properties. We report results of a survey with 2500 US-based participants, which was assessed using structural equation modelling techniques, including factor and path analysis. The resulting model consists of 11 latent factors, which form a three-layer structure. Two predictor layers include DA (users' attitudes and familiarity with them) and task factors (need for control/transparency, subjectivity, risk, self-efficacy, and frequency). Their influence on willingness to delegate is mediated by a values layer (trust, perceived ease of use, and usefulness). This model exhibited an adequate fit (2 = 3529, df = 827, CFI = 0.947, TLI = 0.940, RMSEA = 0.043, SRMR = 0.098). Cronbach's alpha and composite reliability values were satisfactory for each factor, falling in range between 0.69 and 0.94. Overall, the developed model demonstrated good psychometric properties. Further analysis of the relationships between the factors implied a mismatch between users' expectations for DAs and their actual experiences. These findings could be useful to guide practitioners who work on design and development of DAs.

Title

Modeling personality language use through semantic vector subspaces

Author(s)

José Ángel Martínez-Huertas[1], Guillermo Jorge-Botana[2], Alejandro Martínez-Mingo[3], José David Moreno[3], Ricardo Olmos[3]

[1]National Distance Education University, Madrid, Spain; [2]Complutense University of Madrid, Madrid, Spain; [3]Autonomous University of Madrid, Madrid, Spain

Abstract

We propose a cost-effective method for generating observable semantic indicators/sensors that capture relevant variability of the Big Five model from language. To do so, we used hierarchical semantic vector subspaces, a new computational development from vector space models. This poster summarizes the validity evidence of semantic vector subspaces from two studies using two different prompted-based self-descriptions. The constructed responses were answered by 643 Spanish native speakers. We used standardized multiple-choice tasks as validity criteria for the computational scores of the hierarchical semantic vector subspaces. In the first study, we found convergent and discriminant validity of different latent profiles of personality language use. In the second study, different linear relations were observed between the language indicators and the personality traits showing the differences between the profiles found in the first study. Results suggest that these hierarchical semantic vector subspaces can be used to extract personality trait-relevant semantic properties from language, which has methodological and theoretical implications for the study of language and personality relations.

Title

Evaluating Technology Enhanced Learning by Using Single-Case Experimental Designs: A Systematic Review

Author(s)

Nadira Dayo, Wim Van den Noortgate

KU Leuven, Leuven, Belgium

Abstract

Single-Case Experimental Designs (SCEDs) may be a reliable and internally valid way to evaluate Technology Enhanced Learning (TEL) due to the repeated measurements, consideration of individual differences and prevention of logistical issues while testing new technology as the small sample allows for a rigorous follow up of the intervention. However, there is no systematic review describing how and why SCEDs were used for TEL evaluation. Therefore, this study conducted a systematic review of the design characteristics, EdTech tools, outcome variables, justifications for use, data collection tools and analysis techniques, and problems and limitations of using SCEDs to evaluate TEL. Accordingly, eight databases were searched and from 2166 hits, data was extracted from 136 studies fulfilling the inclusion criteria. Results indicated that multiple baseline designs were mostly used to evaluate TEL, and intervention phase was repeatedly measured. The extensively-used EdTech tools and techniques were: CAI, video modelling, digital games, mobile applications, virtual and augmented reality and these were typically used to improve and develop language, social behavior, daily living tasks, mathematical concepts and skills. The frequent used data collection tools were: observations, tests, quizzes, questionnaires and task analyses with visual analysis being the widely used data analysis technique. 76 studies did not acknowledge any problem/limitation, while some studies reported small sample and generalization as limitations of using SCED for TEL. The study provides valuable information to utilize SCEDs to advance TEL evaluation methodology. The study ends with a reflection on further opportunities SCEDs can offer for evaluating TEL.

Title

Development and validation of a measure of physicians' erroneous assumptions toward intellectual disability

Author(s)

Alice Bacherini[1], Pasquale Anselmi[2], Susan Havercamp[3], Giulia Balboni[1]

[1]University of Perugia, Perugia, Italy; [2]University of Padua, Padua, Italy; [3]The Ohio State University, Columbus, USA

Abstract

An up-to-date, valid, and reliable measure of physicians' erroneous assumptions toward intellectual disability (ID) is currently unavailable, despite the importance of this construct for medical disability education.
This contribution presents the development and validation of a new measure of physicians' erroneous assumptions toward ID.
First, the authors wrote a list of 100 possible physicians' erroneous assumptions toward ID and refined them through two field tests with four ID experts and a social psychologist. Then, 133 American ID stakeholders rated each assumption on two 5-point Likert scales, one evaluating the perceived prevalence of the assumption in physicians and the other evaluating its damage for the health care of adults with ID. An assumption was considered prevalent in physicians and damaging for the healthcare of adults with ID if rated as held by "A lot" or "Most" of physicians and as "Significantly damaging" or "Very damaging", respectively. Thus, the most prevalent and most damaging erroneous assumptions were selected (n = 27) and their prevalence and damage ratings were ascertained to be unidimensional, reliable, and independent from the stakeholders' characteristics.
These 27 erroneous assumptions comprised a new scale to measure this construct in physicians, where participants are asked to express their level of disagreement/agreement (5-point Likert scale) with each statement. A validation study was conducted on data collected on 279 American physicians. The instrument showed strong psychometric properties, verified using both classical test theory (factor analysis, convergent/divergent validity, reliability) and item response theory (item location, discrimination, DIF).

Title

Controlling the Carry-Over Effect across Different Scales in Moderation Analyses

Author(s)

Yi-Jhen Wu[1], Kuan-Yu Jin[2], Chia-Ling Hsu[2], Yi-Hsin Chen[3]

[1]The Center for Research on Education and School Development, TU Dortmund, Dortmund, Germany; [2]Hong Kong Examinations and Assessment Authority, Hong Kong, Hong Kong; [3]College of Education, University of South Florida, Tampa, USA

Abstract

Many social science surveys are designed to measure different but correlated constructs using the one-statement-multiple-scale (OSMS) format. The OSMS format is that the same statement for each item will be presented in different constructs (e.g., the perception of the frequency of pain and the perception of the severity of pain). However, this format may lead to the carry-over effect that responses to a previous construct affect responses to a following construct. This violates the local independence assumption in most measurement models and may bias associations between latent variables. In this study, we aimed to investigate the consequences of model misspecification for an OSMS format in moderation analyses via simulations and real examples. To address how ignoring the carry-over effect in the OSMS format, we compared two different measurment models in moderation analyses using simulation and real data: (1) a conventional multidimensional IRT model in a measurement part (Model 1), and (2) an advanced multidimensional IRT model for the carry-over effect in a measurement part (Model 2). Our results showed that ignoring carry-over effects (Model 1) led to biased regression coefficients in moderation analyses. In contrast, a multidimensional IRT model for the carry-over effect (Model 2) yielded more accurate estimates of moderation effects. These findings have important implications for appropriately addressing carry-over effects in OSMS format surveys and moderation analyses.

Title

The Relationship between LIWC Linguistic Indicators and Personality: A Failed Cross-Validation Study

Author(s)

José David Moreno[1], José Ángel Martínez-Huertas[2], Guillermo Jorge-Botana[3], Alejandro Martínez-Mingo[1], Ricardo Olmos[1]

[1]Universidad Autónoma de Madrid (UAM), Madrid, Spain; [2]Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain; [3]Universidad Complutense de Madrid (UCM), Madrid, Spain

Abstract

Background: An emerging field of research points towards the idea that there are personality characteristics that can be reflected in the language that people use. Previous meta-analytic research found small to moderate relationships between the Big Five personality traits and different linguistic computational indicators. However, previous studies included multiple linguistic indicators to predict personality from an exploratory framework. The aim of this study was to conduct a cross-validation study analyzing the relationships between language indicators and personality traits to test the generalizability of previous results; Methods: 643 Spanish undergraduate students were tasked to write a self-description in 500 words and to answer a standardized Big Five questionnaire. The LIWC tool was used to evaluate multiple linguistic indicators from the self-descriptions of the participants. Two different analytical approaches using multiple linear regression were followed: First, using the complete data and, second, by conducting different cross-validation studies; Results: The results showed moderate effect sizes and significant relationships between language and personality in specific samples, but the cross-validation study showed that the model estimates were not generalizable to other samples; Conclusions: Moderate effect sizes were obtained when the language and personality relationships were analyzed in single samples, but it was not possible to generalize the model estimates to other samples. Thus, previous exploratory results found on this line of research appear to be incompatible with a nomothetic approach. Although this is a promising research field, we should emphasize the need for robust cross-validation methods in the analysis of predictive indicators of personality traits.

Title

Identifying item misfit in cognitive diagnostic modeling with small sample sizes

Author(s)

Miguel A. Sorrel[1], Rodrigo S. Kreitchmann[2,3], Pablo Nájera[1,2], Francisco J. Abad[1]

[1]Universidad Autónoma de Madrid, Madrid, Spain; [2]Universidad Alfonso X el Sabio, Madrid, Spain; [3]Instituto de Empresa, Madrid, Spain

Abstract

Cognitive diagnostic models are confirmatory latent class models that classify examinees on a set of discrete latent dimensions. Their main area of application is in educational measurement, where the diagnostic output these models provide can be used to guide teaching efforts according to the students' strengths and weaknesses. However, in these contexts, the small sample size compromises the estimation of item parameters, disrupting the classifications and analyses that rely on these parameters, such as the assessment of fit. One proposed solution is using simpler models to facilitate parameter estimation. The aim of this study is to test the performance of some fit statistics available in the literature, as well as a proposal based on Stone's method that considers classification uncertainty and another one inspired by the nonparametric classification method, which does not rely on parameter estimates. For this purpose, a Monte Carlo simulation study was designed in which, among other factors, the sample size and the quality of the items were manipulated. The performance of the different alternatives is assessed in terms of Type I error (% of items for which the true, data-generating model is fitted, but the method indicates misfit) and Power (% of items for which a false, non-data-generating model is fitted, and the method indicates misfit). The results are promising, especially for the two original proposals of the study. An empirical example of how to implement the proposed methods with R is described.

Title

Comparing the Pace of Psychological Change in Videoconferencing (VCP) and Face-to-Face (F2F) Psychotherapy: A Longitudinal Multilevel Growth Curve Modelling Approach.

Author(s)

Diego Fernández-Regueras, M. Cristina Guerrero-Escagedo, Alba Luque-González, Ana Calero-Elvira

Universidad Autónoma de Madrid, Madrid, Spain

Abstract

The purpose of this study is to explore the pace of psychological change in face-to-face (F2F) and videoconferencing psychotherapy (VCP) and to offer a methodological tool for studying it. Additionally, the study aims to generate hypotheses that could explain the differences between the pace of change in F2F and VCP. We hypothesized that change in therapy would be non-linear and faster in F2F than in VCP. We collected session-by-session records of two measures of change in therapy (as assessed by therapists and clients) from 113 participants in F2F (n = 57) and VCP (n = 56). This resulted in a sample of 2552 therapy sessions. We proposed a non-manipulative longitudinal design that used multilevel growth curve modelling. Different models were adjusted to account for the trajectories followed by all cases as closely as possible. The chosen models, adjusted for therapists' and clients' data, showed medium and large effect sizes. The results indicated that change was indeed non-linear and faster in F2F, as we had predicted. We have generated several hypotheses that attempt to explain the processes that could be behind these results, particularly those related to the therapeutic alliance. We call for future studies to continue investigating this research area, while also working to refine the methodology necessary for studying it.

Title

Development and Reliability of an Observational Methodology Instrument for the Assessment of the Therapeutic Relationship.

Author(s)

M. Cristina Guerrero-Escagedo, Diego Fernández-Regueras, Alba Luque-González, Ana Calero-Elvira

Universidad Autónoma de Madrid, Madrid, Spain

Abstract

The therapeutic relationship plays a crucial role in psychological therapy. However, many studies rely on self-report measures rather than external observation to identify the behaviours that enhance it. To address this gap, we aimed to develop an observational coding system for individual therapy with adults using a cognitive-behavioural approach. In this communication, we present the process of constructing and validating this instrument, following Bakeman & Quera's (2011) guidelines. Specifically, we followed six steps: (1) formulated a research question, "Which behaviours are relevant to the therapeutic relationship?" and conducted a literature review to identify factors for observation; (2) established a social criterion as the level of analysis, which enabled us to assess behaviours whose distinctions were not purely physical; (3) established observation conditions; (4) created the final categories for our coding system, which has two sub-systems: the therapist's verbal behavior and the client's verbal behavior; (5) observed 50 video recordings of individual psychotherapy sessions and used team discussions to refine the coding system until adequate Cohen's Kappa levels were achieved in both sub-systems (therapists .63-.75; clients .60-.75); and (6) made changes based on our observations. Our coding system enables moment-to-moment analysis of the therapeutic relationship during sessions, identifying the elements that contribute to successful outcomes. It represents the first step towards systematically analyzing the therapeutic relationship through therapist-client interaction and can contribute to future studies as recommended by the APA's Division 29 Task Force. Additionally, we aim to expand methodologists' knowledge about the development and usage of observational methodology instruments.

Title

Bullshit susceptibility scale: Development and evidence of validity in adult population

Author(s)

Geraldy Sepúlveda[1], Bárbara Rodríguez[2]

[1]Universidad de Tarapacá, Arica, Chile; [2]Universidad de Talca, Talca, Chile

Abstract

Research on disinformation has acquired great relevance in the era of social media given the massive social, sanitary and political impacts that it has. In this context, fake information without any specific purpose or bullshit is being spread with ease in social media. It is relevant to identify people's characteristics that could moderate the credibility that they give to disinformation. Susceptibility/receptivity to bullshit has evidence to be a relevant factor. However, despite the interest in this construct, the available instruments are scarce and generated from experimental expositions or ad hoc scales without enough psychometric support which limitate the development of research of this construct. Hence, the purpose of this study was to develop a brief scale to be incorporated in large studies, containing evidence of reliability and validity to measure susceptibility to bullshit in the adult population. We conducted a psychometric study in general population, with a preliminary exploratory phase (n=118) and confirmatory (n=450) using ESEM. The final instrument is constituted of two dimensions (Sense and Nonsense) and 22 items and adequate levels of reliability (>.80). Furthermore, it has evidence of internal structure validity through ESEM (CFI .95 ; TLI .95; RMSEA < .60), invariance between men and women, and evidence of validity related to other variables (age, religion, ethnicity, alternative medicine practices, paranormal beliefs, cognitive abilities). Finally, we discussed the reaches and limitations of this instrument, as well as its possible application in social and health psychology.

Title

A Network Analysis Approach to Explore the Interrelationships Among Democratic Competences

Author(s)

Giusy Danila Valenti, Nicolò Maria Iannello, Maria Valentina Cavarretta, Alida Lo Coco, Cristiano Inguglia, Sonia Ingoglia

Department of Psychology, Educational Science and Human Movement, University of Palermo, Palermo, Italy

Abstract

Network Analysis (NA) may be useful to estimate the reciprocal influence among a set of variables. To date, no existing studies have been conducted in which this analytic approach has been applied to investigate the relationships among democratic competences (i.e., empathy, respect, responsibility, and cooperation) and civic engagement (attitudes and behaviors). We aimed to conduct NA to examine the interrelations among the aforementioned variables.
We recruited a sample of 441 adolescents (70.1% females; Mage = 16. 51, SD = 1.36), from some High Schools in Southern Italy. A graphical least absolute shrinkage and selection operator (LASSO) regularization based on the Extended Bayesian Information Criterion (EBIC) was used, with a tuning parameter ( ) of .50. We inspected centrality indices (i.e., strength, betweenness, and closeness), as well as edge stability by the bootstrapping procedure (1000 resampling) at 95% CI.
We found 12 positive non-zero edges, with civic attitudes and behaviors showing the strongest associations (r = .32). Responsibility was not directly connected to neither civic attitudes nor civic behaviors, whereas respect did not report direct associations with civic behaviors. Cooperation and civic engagement (attitudes) were the most central nodes, representing the nodes with the highest overall influence in the network, as well as indicating that they functioned as a bridge between other nodes. Stability analysis revealed that the network was accurately estimated, with moderate CI around the edge weights.
Our study offers a contribution to the understanding of some democratic competences during adolescence, also pointing out on their mutual interrelationships.

Title

Development of a Revised Short Dark Triad Questionnaire Using Ant Colony Optimization Algorithms.

Author(s)

Lukas A. Knitter[1], Alisa Lange[1], Maximilian Kluge[1], Martin Schultze[2], Tobias Koch[1]

[1]Institute of Psychology, Friedrich Schiller University, Jena, Germany; [2]Institute of Psychology, Goethe University, Frankfurt, Germany

Abstract

The Dark Triad of personality consists of three malevolent traits: Machiavellianism, psychopathy, and narcissism. Researchers have developed several instruments to measure these traits individually. However, most of them are lengthy and time-consuming, which limits their usefulness in research settings. Short scales for measuring the three traits simultaneously have been criticised in the literature. These measures have been shown to have problems with construct validity. Many authors have attempted to compensate for this weakness by using different measurement models, e.g., bifactor models. The results of a recent meta-analysis (Knitter et al., 2023) showed that different measurement models do not address the psychometric weaknesses of the original scale. In particular, the original Short Dark Triad questionnaire lacks discriminant validity between the three traits (facets), especially between Machiavellianism and psychopathy.

In this talk we present a revised version of the Short Dark Triad questionnaire. In the revision, we reformulated and added items to improve discriminant validity and to address criticisms in the literature (Miller et al., 2019). We collected data from two samples, along with external criterion variables, to assess the quality of the new items. We used the R package stuart (Schultze, 2022) for criterion-guided and automated item selection, prioritising selected items that maximised reliability within each facet, maximised discriminant validity between facets, and met theoretical expectations regarding relationships with multiple criterion variables (i.e., personality dimensions, life satisfaction, and political attitudes). In addition, expert judgement was used to assess the content validity (wording) of the items in relation to the respective traits.

Title

Utilizing the mind map technique to identify indicators of attitudes towards scientific research among teacher candidates

Author(s)

Elisabeth Desiana Mayasari[1,2]

[1]University of Łódź, Łódź, Poland; [2]Sanata Dharma University, Yogyakarta, Indonesia

Abstract

Developing attitudes toward scientific research is crucial for teacher candidates as it helps them cultivate critical and reflective attitudes toward their future teaching practices. These attitudes enable them to enhance their knowledge by assessing both their successes and areas that require improvement in their teaching methods. However, students tend to avoid research-related classes due to anxiety and the perception that they pose a barrier to their studies. Therefore, the objective of this study is to identify the factors that contribute to the formation of attitudes toward scientific research among teacher candidates.
The study utilized a qualitative method, specifically interviews, as the research instrument. The researcher employed a mind map for the qualitative data analysis, which involved conducting semi-structured interviews, transcribing the interviews, creating a mind map of the interview outcomes related to the study's topic, and discussing the mind map results with the participants to validate and gather more data. Subsequently, the researcher collected the themes that emerged from the mind map as the basis for writing the study's findings. The study involved participants from Iran, Turkey, and Indonesia.
The findings of the study indicate that various factors contribute to the formation of attitudes toward scientific research in teacher candidates. These factors include advantages, skill acquisition, capability, emotions towards research, environmental factors, and psychological factors. These findings can provide insights for educators to develop effective strategies to promote positive attitudes toward scientific research among teacher candidates.

Title

Comparing Equally and Unequally Keyed Forced-Choice Questionnaires for Measuring Personality Traits

Author(s)

Diego F. Graña[1], Rodrigo S. Kreitchmann[2], Francisco J. Abad[1], Miguel A. Sorrel[1]

[1]Universidad Autónoma de Madrid, Madrid, Spain; [2]IE University, Madrid, Spain

Abstract

Interest in measuring non-cognitive traits has increased in recent decades revealing problems related to its validity, like response biases such as acquiescence or socially desirable responding. To address these biases, forced-choice (FC) questionnaires have been proposed, as they should solve them by design. While some challenges associated with this format have been addressed (e.g. ipsativity can be partially solved with an adequate modeling), others remain as an open scientific debate. One such debate concerns the inclusion of unequally keyed items, with evidence supporting both their inclusion and avoidance. To contribute to this debate, we collected data from 1,125 undergraduate Psychology students who completed a personality item pool measuring the Big Five personality traits in Likert-type format and two FC questionnaires. The FC questionnaires were assembled from the item pool through an optimization algorithm, with one questionnaire consisting only of equally keyed items and another including unequally keyed items. We also included two reference Big Five personality questionnaires for convergent validity and different variables for criterion validity. After IRT calibration, we compared the questionnaires in terms of reliability, convergent and criterion validity, and ipsativity. Our statistical analysis showed isolated differences between the equally and unequally keyed questionnaires, but no systematic differences, indicating that neither questionnaire outperformed the other. We conclude that, for optimally designed tests, this comparable results across formats suggests a preference for equally keyed blocks since unequally keyed ones are more difficult to pair in social desirability, which could theoretically result in vulnerability to response biases.

Title

Methodological preferences - systematic literature review

Author(s)

Martyna Jarota[1,2]

[1]University of Lodz Doctoral School of Social Sciences, Lodz, Poland; [2]Faculty of Educational Sciences of the University of Lodz, Lodz, Poland

Abstract

During planning and preparing the research project, as well as during its implementation, the researcher faces the need to make multiple methodological choices, including theories and methods of data collection and analysis. The researcher's decisions are regulated in the form of specific rules and procedures that make up the theory of scientific methods. It should be noted that important factors guiding scientific research are also methodological preferences of the researcher which influence his substantive research choices. In the discourse surrounding research undertaken within a given scientific discipline, it is important not only WHAT researchers do (research problem), but also HOW they do it. It happens that issues related to research methodology are in fact are elements that determine the direction of the researcher's search, which in fact suggests that methodological preferences play a significant role in organizing knowledge in the field of particular scientific disciplines. The aim of the systematic literature review was to determine the current state of knowledge about methodological preferences, what has been done. Moreover, a gap regarding methodological preferences has been identified in the recent scientific achievements.

Title

Research project as a decision-making process - reflections on the research process in the light of decision theory

Author(s)

Martyna Jarota[1,2]

[1]University of Lodz Doctoral School of Social Sciences, Lodz, Poland; [2]Faculty of Educational Sciences of the University of Lodz, Lodz, Poland

Abstract

The analysis of the research process in the social sciences allows us to see the numerous stages that make them up. At each of the stages, the researcher undertakes many complex activities - this applies to both the conceptual and implementation phases of research. These activities require the researcher to make numerous choices that must be justified. For this reason, the research process can be interpreted as a series of decisions that require the researcher to provide appropriate arguments, e.g. methodological. Therefore, it is reasonable to consider the research process in the light of decision theory. These considerations may constitute an element of methodological criticism as well as methodological reflection. Both (criticism and reflection) can contribute to the development of the methodology of individual scientific disciplines. The purpose of reflecting on a research project as a decision-making process is to better understand the choices that researchers make. What's more, thanks to the reference to decision theory, new contexts and conditions in which research can be carried out have emerged.

Title

Naive skepticism scale: development and evidence of validity

Author(s)

Rodrigo Ferrer[1], Yasna Ramirez[1], Camila Butt[1], Patricio Mena[2], Geraldy Sepúlveda[1]

[1]Universidad de Tarapacá, Arica, Chile; [2]Universidad de la Frontera, Temuco, Chile

Abstract

Traditionally, skepticism has been associated with critical thinking; however, philosophy has proposed the existence of a particular type of skepticism, naive skepticism, which could make people more vulnerable to misinformation as opposed to information from official sources. Although some scales have been proposed to measure skepticism in specific topics, there are no instruments available in the literature to assess this construct. Therefore, the aim of this study was to develop a scale to measure naïve skepticism, through 2 samples in an adult population: a pilot study (n = 126) and a validity evidence study (n = 320). The final scale was composed of 14 items and 2 dimensions (skepticism towards governmental organizations and official press; and skepticism towards science). The results show that the identified structure provides adequate levels of reliability (> 0.8), evidence of validity, based on the internal structure of the test, through ESEM (CFI = .966; TLI = .951; RMSEA = .079), as well as evidence of validity based on the relationship with other variables. Finally, it is concluded that the scale developed has sufficient psychometric properties to interpret the scores as representations of naïve skepticism, and some limitations and possible implications of the scale and the theoretical construct are discussed.

Title

ptchart : An R package for computing Precision Teaching measures and charts

Author(s)

Alexandre Gellen-Kamel[1], Pier-Olivier Caron[2]

[1]Université du Québec à Montréal, Montréal, Canada; [2]Université TÉLUQ, Montréal, Canada

Abstract

Precision Teaching is a domain within Applied Behavior Analysis that are interested in measuring and charting behavior on a standardized chart to assess the effectiveness of an intervention and, if needed, to make the appropriate changes (Heron, Heward & Cooper, 2014).
All calculations and chartings are traditionally made on paper and pencil with the Standard Celeration Chart and with a celeration and frequency finder (Graf & Lindsley, 2002; Pennypacker, Gutierrez & Lindsley, 2003). Even though the paper and pencil approach is suitable for applied purposes, it is not for more rigorous, scientific ones. Very few digital tools exist and most are designed for clinic purposes. There is no existing software tool to help researchers in the Precision Teaching field.
As there is a growing use of the R language among academics and in psychology in general, a dedicated package would be a substantial contribution in Precision Teaching. Thus, we developed 'ptchart', a package that computes measures and produces charts related to Precision Teaching. The goals of 'ptchart' are multiple: (a) provide an intuitive interface with useful default parameters, (b) reduce coding time to get results, (c) give flexibility to manipulate outputs for further analyses, (d) generate charts that are presentation or publication ready. The 'ptchart' package provides two main functions : 'ptstat()' which computes relevant indices, and 'ptchart()' which produces the charts. A thorough example of the package will be presented. As it is an active package in current development, present and planned features and limitations will be discussed.

Title

Measuring Recovery in Spanish-Speaking Population with Severe Mental Disorders: Validation of the Maryland Assessment of Recovery Scale (MARS-12)

Author(s)

Jone Aliri[1], Nekane Balluerka[1], Arantxa Gorostiaga[1], Hernán Sampietro[2], Ana González-Pinto[3]

[1]University of the Basque Country UPV/EHU, Donostia, Spain; [2]ActivaMent Catalunya Associació, Barcelona, Spain; [3]University of the Basque Country UPV/EHU, Gasteiz, Spain

Abstract

This study aimed to culturally adapt and evaluate the psychometric properties of the abbreviated version of the Maryland Assessment of Recovery Scale (MARS-12) in a Spanish sample of individuals with severe mental disorders. Data were collected between January and October 2022, and the study comprised two phases. In Phase 1, a standardized method of translation and back-translation was used to ensure semantic, linguistic, and contextual equivalence between the adapted and original versions of the scale. In Phase 2, the psychometric properties of the Spanish version were evaluated in a sample of 325 individuals with severe mental disorders. The confirmatory factor analysis confirmed that the optimal structure of the MARS-12 was the one-dimensional structure observed in the original scale and demonstrated adequate internal consistency. The total score of the MARS-12-ESP showed a high correlation with the score of the Questionnaire on the Recovery Process, indicating good convergent validity. Furthermore, the correlation of the total score of the MARS-12-ESP with both the score of the Spanish version of the Dispositional Hope Scale and the Multidimensional Scale of Perceived Social Support provided evidence of validity based on the relationship of recovery with dispositional hope and perceived social support. Overall, this study contributes to the development of a reliable and valid instrument for assessing recovery in Spanish-speaking individuals with severe mental disorders.

Title

Validation of the Spanish version of the Game Transfer Phenomena Scale – Short Form: A preliminary study

Author(s)

Laura Maldonado-Murciano[1,2,3], Angelica Ortiz de Gortari[4], Maite Barrios[2,3], Juana Gómez-Benito[2,3], Georgina Guilera[2,3]

[1]Center of Excellence in Responsible Gaming, University of Gibraltar, Gibraltar, Gibraltar; [2]Faculty of Psychology, University of Barcelona, Barcelona, Spain; [3]Institute of Neurosciences, University of Barcelona, Barcelona, Spain; [4]Centre for the Science of Learning & Technology, University of Bergen, Bergen, Norway

Abstract

Game Transfer Phenomena (GTP) refer to the transfer of video game experiences into real life (i.e., altered sensory perceptions, mental processes and behaviours) (Ortiz de Gortari, 2019). The shortest tool to assess the GTP is the short Gaming Transfer Phenomena Scale (GTPS5-SF) (Ortiz de Gortari, Diseth, Styvertsen, & Ståle, 2023), with 5 items measuring altered perceptions in different sensory channels, automatic thoughts and behaviours/actions, based in the original he GTP Scale (GTPS) (Ortiz de Gortari, Pontes, & Griffiths, 2015) with 20 items. The objective of this study was to preliminary validate the Spanish version of the GTP5-SF. A sample of 120 gamers (51.67% women, mean age 25.83 years, SD = 9.87) participated. The GTP5-SF was adapted from English to the Spanish language using parallel translation. Item descriptives were obtained. Unidimensionality of the GTP5-SF was tested using confirmatory factor analysis. Internal consistency of scores was assessed computing Cronbach's alpha and McDonalds' omega. Finally, GTP5-SF score was correlated with gaming disorder. Analyses were carried out with the R packages lavaan and psych. Item response distribution appeared to be right-skewed. The one-factor structure was confirmed (CFI = 1, SRMR = .020), with item loadings ranging from .78 to .96. Cronbach's alpha and omega coefficients reached values of .86 and .88, respectively. The GTP5-SF score strongly correlated with measures of problematic gaming and session length. GTP prevalence was 46.67%. The preliminary assessment of the psychometric properties of the Spanish version of the GTP5-SF shows that it is a suitable tool for measuring GTP.

Title

Development and content validation of the Facilitators and Obstacles of Recovery Scale (FOR-S)

Author(s)

Estefania Guerrero[1], Hernán Sampietro[2], Maite Barrios[1], Georgina Guilera[1], Juana Gómez-Benito[1]

[1]Universitat de Barcelona, Barcelona, Spain; [2]ActivaMent Catalunya Associació, Barcelona, Spain

Abstract

Nowadays, public mental health policies prioritize recovery-oriented approaches, but there is a lack of validated scales that consider the facilitators and barriers of the recovery process. This study aimed to develop and validate the content of a psychometric instrument to assess obstacles and facilitators of mental health recovery. Identifying available facilitators promotes positive support and enables focusing on the users' strengths; meanwhile, recognizing the presence of the obstacles helps to engage necessary efforts to overcome or reduce them.
Items were developed from themes agreed upon in a previous Delphi study with 81 members of users and survivors of psychiatry organizations. The results of this Delphi study identified 12 main themes related to obstacles and eight main themes related to facilitators of recovery. A preliminary version of the Facilitators and Obstacles of Recovery Scale (FOR-S) was then presented to a panel of 20 experts on a recovery-oriented approach to establish content validity. The panel reviewed the themes and scored the items using a relevance Likert-type scale. The Item-level and Scale-level Content Validity Index were calculated to get the final set of items. In the second phase of the study, ten users of mental health services evaluated the intelligibility of the instructions and statements using a Likert-type scale. The results of this study led to the development of a 20-item scale for assessing obstacles (12 items) and facilitators (8 items) in the recovery process. The FOR-S seems to be a promising self-administrated scale that can be used in clinical and non-clinical settings.

Title

Spanish adaptation of the Self-Identified Stage of Recovery: a preliminary study

Author(s)

Hernán Sampietro[1], Maite Barrios[2], Ángela Berrío[3], Georgina Guilera[3], Juana Gómez-Benito[3]

[1]ActivaMent Catalunya Associació, Barcelona, Spain; [2]Univesity of Barcelona, Barcelona, Spain; [3]University of Barcelona, Barcelona, Spain

Abstract

The Self-Identified Stage of Recovery (SISR) is a two-part scale. The SISR-A is a one-item forced choice sub-scale that evaluates the self-perceived stage of recovery, and the SISR-B assesses the four key components of the recovery process: hope, identity, meaning, and responsibility. Some studies have provided evidence of its psychometric validity and reliability in the English and Japanese versions, but there is no Spanish version available. The aim of this work was to adapt the English version of the SISR into Spanish and to provide preliminary evidence of its reliability and validity. First, four independent translators conducted a forward translation, followed by an evaluation on the comprehensibility of the items by a panel of 10 experts by experience using a 1-4 rating scale. Items with a score below 4 were revised, and a subsequent forward translation was conducted based on the suggested changes. Secondly, 120 users of mental health services were recruited from 14 community mental health services in Catalonia, Spain. We evaluated internal structure using CFA, internal consistency with the McDonald's , temporal stability by ICC, and relationships with other variables through the Spearman's rho. We found evidence to support a one-factor structure of the SISR-B. The SISR-B had good internal consistency (.80) and a strong relationship with MARS12 (.74). Both SISR-A and SISR-B have shown adequate temporal stability (.63 and .78, respectively). These preliminary results support the reliability and validity of the Spanish version of the SISR among users of mental health services in Spain.

Title

A proposal for designing interview protocols to explore sensitive information: An application in making-decisions processes related to moral dilemmas.

Author(s)

Belén Carrascal-Caputto[1,2], Araceli Méndez-Andrade[3], Isabel Benítez[1,2], Pilar Aguilar[3]

[1]University of Granada, Granada, Spain; [2]Mind, Brain and Behaviour Research Center (CIMCYC), Granada, Spain; [3]Universidad Loyola Andalucía, Sevilla, Spain

Abstract

Interviews have been frequently used to deeply explore the lived experiences of participants in relation to a specific phenomenon. However, sometimes the phenomenon of interest is related to difficult situations involving sensitive and private information, what could lead to difficulties reaching the intended goals. An example of this would be decision-making in the face of moral dilemmas, as moral dilemmas usually reflect situations related to uncomfortable topics. The aim of this study is to identify the most adequate probes for capturing emotions and cognitive processes through semi-structured interviews while avoiding both causing discomfort to participants and threatening their privacy. To achieve this goal, participants were first exposed to a battery of sacrificial moral dilemmas. Subsequently, they were provided with the definition of the characteristics of a dilemmatic situation in real life. With the definition in mind, they were asked to recall in their memory a moral dilemma they had experienced in their real lives. Finally, they answered a series of questions around that real experience related to the emotions and cognitive decision-making processes they carried out in the face of the moral dilemma. Results provided relevant information about the psychological processes experienced in these dilemmatic situations which could be considered equivalent to those theoretically expected but without the need for the participants to explicitly describe the situation or details about the decisions they made. Details about the most efficient probes and practices will be provided, as well as suggestions for designing interview protocol focused on private or sensitive topics.

Title

Exploring approaches for estimating parameters in cognitive diagnosis models with small sample sizes

Author(s)

Miguel A. Sorrel[1], Scarlett Escudero[2], Pablo Nájera[3,4], Rodrigo S. Kreitchmann[4,5], Ramsés Vázquez-Lira[2]

[1]Madrid, Madrid, Spain; [2]Universidad Nacional Autónoma de México, Ciudad de México, Mexico; [3]Universidad Autónoma de Madrid, Madrid, Spain; [4]Universidad Alfonso X el Sabio, Madrid, Spain; [5]Instituto de Empresa, Madrid, Spain

Abstract

In recent years, cognitive diagnostic models (CDMs) have gained increasing popularity in various assessment contexts, with the DINA model being the most widely adopted. CDMs are designed to diagnose cognitive processes underlying an examinee's performance, making it possible to provide tailored feedback. However, the most commonly used parameter estimation method for CDMs, marginal maximum likelihood using the Expectation-Maximization algorithm (EM-MMLE), can present difficulties when sample sizes are small. This study aims to compare the results of different estimation methods for CDMs under varying sample sizes. Specifically, we compare EM-MMLE, Bayes modal, Markov chain Monte Carlo (MCMC) with Gibbs sampler, MCMC with Hamiltonian Monte Carlo sampler, a non-parametric method, and a parsimonious parametric model such as R-DINA. We use both simulated and empirical data, varying the sample size from small to large, and assess the bias in the estimation of item parameters, the precision in attribute classification, the bias in the reliability estimate, and computational cost. Our findings suggest that all other options are preferred over EM-MMLE under conditions of low sample size, whereas comparable results are obtained under conditions of large sample size. Thus, practitioners should consider using alternative estimation methods when working with small samples to obtain more accurate estimates of CDM parameters. By providing guidance on the estimation of CDM parameters, this study aims to maximize the potential of CDMs for improving educational practice.

1.17 Panel discussion 17h30–18h30

Title

Career Discussion: Experiences Working Outside Academia

Author(s)

Miles Jeremy[1], Sengewald Erik[2]

[1]Google, Los Angeles, USA; [2]German Federal Employment Agency, Fürth, Germany

Abstract

People who have experience and skills in quantitative social science may find that their expertise is in demand in a wide range of industries and areas of application. This symposium will bring together a number of people who apply quantitative social science methodology in a range of areas, to demonstrate the types of different roles that social scientists may play outside of traditional academic research.
Each speaker will briefly introduce themselves and describe their career path. For the remainder of the session we will open the floor to questions, which can be asked live, or in advance, using the website tinyurl.com/CareersDory (you may ask questions anonymously, or under your name).
Speakers:
Erik Sengewald, Senior expert for test development, Psychological Service - Research and Development, German Federal Employment Agency (PhD Psychology, University Jena, Germany)
Jeremy Miles, Data Scientist, Google, USA (PhD Psychology, University of Derby, UK)
Amelie Vrijdags, Senior Consultant - Expert Psychologist, Hudson Benelux (PhD Psychology, Ghent University, Belgium)
Jonas Tundo, co-founder and CEO, Dataroots (Master after Master Statistical Data Analysis, Ghent University, Belgium)
Maarten De Schryver, Senior Manager - People Analytics, Deloitte (PhD Psychology, Ghent University, Belgium)
Han Bossier, Statistical Consultant, OpenAnalytics (PhD Psychology, Ghent University, Belgium)
Chair: Steffi Pohl, Department of Education and Psychology, Freie Universität, Berlin.

2 Wednesday 12 July

2.1 Keynote speaker 10h00–11h00

Title

Sample size calculations

Author

Mirjam Moerbeek

Utrecht University

Abstract

noindent One of the main steps to be taken in the design of a study is the calculation of sample size. In this presentation I will give a summary of my past, present and future research on this topic, with a focus on cluster randomized trials. With cluster randomized trials, complete clusters such as schools, general practices or neighborhoods are randomized to treatment conditions and all subjects in the same cluster receive the same condition.

The first part of this presentation focused on sample size calculations from a frequentist point of view. It will be shown how to calculate how many clusters and how many subjects per cluster should be included in the trial. These sample sizes can be shown to depend on the intra-class correlation coefficient. An a priori estimate of this model parameter is not always available and various approaches to deal with this will be discussed.

The second part of this presentation focuses on Bayesian sample size calculation. It will be shown how the Bayes factor is used to evaluate informative hypotheses and a criterion for a priori sample size determination is introduced. Furthermore, Bayesian sequential designs are discussed. With such designs, additional subjects are recruited during the course of the study until sufficient support for either informative hypothesis is achieved.

2.2 State-of-the-art 15h00–15h30 Aud 1

Title

Regression models in Causal inference

Author

Rhian Daniel

Cardiff University

Abstract

Although problems in causal inference typically have one relatively simple causal estimand as the ultimate target, parametric regression models (with multiple parameters) are often used as nuisance models in the estimation procedure. This can lead to a number of challenges, such as undiagnosed extrapolation and so-called null paradoxes, the latter being a problem when the regression models include non-collapsible parameters. In this talk, I will give a summary of these challenges, and discuss how a new class of parametric regression models —known as Regression by Composition— may offer some advantages.

2.3 State-of-the-art 15h00–15h30 Aud 2

Title

Finding clusterwise measurement invariance with mixture multigroup factor analysis

Author

Kim De Roover

KU Leuven

Abstract

Psychological research often builds on between-group comparisons of (measurements of) latent variables, for instance, to evaluate cross-cultural differences in mindfulness. A critical assumption in such comparative research is that the same latent variable(s) are measured in the same way across all groups (i.e., measurement invariance). Nowadays, measurement invariance is often tested across lots of groups. When (a certain level of) measurement invariance is untenable across many groups, it is hard to unravel invariances from non-invariances and for which groups they apply. Mixture multigroup factor analysis (MMG-FA; De Roover, 2021; De Roover, Vermunt, & Ceulemans, 2020) was recently proposed to cluster groups based on the measurement parameters, whereas the structural parameters are allowed to differ between groups within a cluster. More specifically, MMG-FA clusters the groups according to a specific level of 'clusterwise measurement invariance' (e.g., based on factor loadings only to achieve metric invariance within clusters, or based on loadings and intercepts to achieve scalar invariance within clusters). In this presentation, the full framework of mixture multigroup factor analysis and the 'mixmgfa' R-package are presented, as well as how to take the steps from the initial overall level of invariance across all groups to the desired level of clusterwise invariance.

2.4 Parallel sessions 08h30–10h00 Auditorium 1

Symposium Overview

Integrating quantitative and qualitative evidence for developing and validating psychological assessments: Challenges and benefits of mixed methodology

Author(s)

Jose-Luis Padilla[1,2], Isabel Benítez[1,2]

[1]University of Granada, Granada, Spain; [2]Mind, Brain and Behaviour Research Center (CIMCyC), Granada, Spain

Abstract

In the last decades, validity theory has addressed new challenges related to emerging administration modes (web, apps, mobile phones, etc.), and social concerns such as values, intended and unintended social consequences, or respondent diversity. These challenges urge professionals involved in psychological and educational assessment to develop a comprehensive validity argument based on an array of validity evidence. Mixed methodology is becoming the preferred methodological approach to perform growing complex validation studies because of the variety of proposals for integrating quantitative and qualitative evidence, among other reasons. The symposium will bring four examples of developing instruments and validation studies to illustrate how integration can be reached through the different stages of the research. The presenters will discuss validity evidence based on response processes associated with survey questions and scales items by cognitive interviewing and web probing, integration of evidence from focus groups and literature review to guide construct definitions, qualitative methods as tool for validation in the increasingly popular Experience Sampling Method, and meaning making in self-report measurement as source of evidence beyond quantitative clinical measurement. With these examples, the panellists encourage the audience to engage in a problem-based thinking exercise on the sources, applications and variation of validity theory, and to discuss the added value of integrating quantitative and qualitative methods for valid psychological assessment.

Title

Developing a questionnaire for assessing "perfectionism" by a mixed-methods approach

Author(s)

Luis Manuel Lozano[1,2], Isabel Benítez[1,2], Andrés González[1], José Luis Padilla[1,2]

[1]University of Granada, Granada, Spain; [2]Mind, Brain and Behaviour Research Center (CIMCYC), Granada, Spain

Abstract

Test development usually follows a series of ordered steps from construct definition to interpretation norms. On the one hand, test developers resort mostly to quantitative techniques to analyze item characteristics, reliability and obtain validity evidence like internal structure, relationships with other variables and criteria. On the other hand, qualitative methods are used to define the intended construct and increasingly for validity sources like response processes or consequences of assessment. The research aims to integrate qualitative and quantitative methods through different phases of developing a questionnaire intended to assess perfectionism in adolescents. In a "multistage mixed methods framework" (Fetters et al., 2013), different approaches to achieve integration at the "design" and "method" levels will be followed. The contribution will share the mixed-methods work done for the definition of perfectionism in Spanish adolescents integrating results from a literature scoping review, content analysis of perfectionism scales available in the literature, and focus groups with adolescents, parents, and teachers. "Building" approach to integration at the method level is used when scoping review and content analysis findings informed the focus group protocols. Subsequent mixed-methods phases in the development of the perfectionism scales will be outlined. Lastly, we will share challenges in developing the scale by a mixed-methods approach to encourage discussion.

Symposium title

Integrating quantitative and qualitative evidence for developing and validating psychological assessments: Challenges and benefits of mixed methodology

Title

Examining how to integrate psychometrics with qualitative evidence from web probing and cognitive interviewing for survey questions and scale items

Author(s)

Jose-Luis Padilla[1,2], Isabel Benítez[1,2], Irene Gómez-Gómez[3], María-del-Carmen Aguilar-Luzón[1,2]

[1]University of Granada, Granada, Spain; [2]Mind, Brain and Behaviour Research Center (CIMCYC), Granada, Spain; [3]Universidad Loyola Andalucía, Sevilla, Spain

Abstract

Since the late 80s of the last century qualitative methods like cognitive interviewing are being used by survey methodologist to research on response processes to survey questions. On the psychometric side, the latest edition of the Standards for Educational and Psychological Testing increased the interest for uncovering the response processes to test and scales items. Researchers and professionals in testing and psychological assessments face the challenge of "combining" common and emerging psychometrics with qualitative findings. The aim of this work is to present and share the work in progress of a research project aimed at developing guidelines and protocols to conduct mixed-methods validation studies. We will present the general objectives and methodologies applied in several mixed-methods studies. First, we will present first results of a validation studies in which Latent Class Analysis results are integrated with qualitative evidence from cognitive interviewing to identify and understand classes of survey respondents to the Patient Health Questionnaire included in the 2020 Spanish Health Interview Survey. Second, we will illustrate how to integrate qualitative findings from web probing with psychometrics to obtain validity evidence of response processes to eco-anxiety scale items and survey questions on environmental topics included in major survey research projects. We will discuss benefits and difficulties in conducting mixed-methods validation studies.

Symposium title

Integrating quantitative and qualitative evidence for developing and validating psychological assessments: Challenges and benefits of mixed methodology

Title

Mixing methods for meaningful measurement in psychological research

Author(s)

Femke Truijens

Erasmus University, Rotterdam, Netherlands

Abstract

Measurement is considered the cornerstone of evidence-based psychology. Mental health research relies predominantly on self-report measures, in which items and responses are standardized to allow for comparison of numbers over people. Researchers often assume that individual differences in scoring are 'covered' by the validity of a measure and seldomly look into scoring processes.
However, scoring self-report measures is inherently meaningful, as it requires people to reflectively translate their experiences into numbers. This presentation features 'John', a patient-participant a gold standard psychotherapy study for major depression treatment. John stood out in the sample, because he was so upset with the research procedure that he actively sabotaged the data collection.
Based on John's case, I argue (1) that self-report measurement is meaningful. In Johns case, his meaning-making process formed a threat to the validity of his data, showing that validity of a measures as such is not sufficient to 'cover' validity of resulting data. Therefore, I argue (2) that active empirical validation within the concrete data collection process is required to go beyond mere face validity. Importantly, though, John's data would not have been identified as invalid by purely looking at the numbers themselves. Rather, it takes his story to identify his particular meaning-making process as threat to validity. Therefore, (3) I argue for mixing methods for validation of self-reported data. I discuss concrete qualitative tools – from open-ended survey questions to cognitive interview and thinking aloud techniques – that allow for meaningful measurement and validation in the action of psychological research.

Symposium title

Integrating quantitative and qualitative evidence for developing and validating psychological assessments: Challenges and benefits of mixed methodology

Title

What idiographic methods can learn from idiographic experiences: The value of mixed methodology for experience sampling studies

Author(s)

Melissa De Smet[1,2]

[1]Ghent university, Ghent, Belgium; [2]Tilburg University, Tilburg, Netherlands

Abstract

While Experience Sampling Methods (ESM) promise important avenues to study person-specific dynamics in daily life, the validity of ESM data and findings pose a challenge. This presentation addresses: 1) the lack of methodological or conceptual guidelines to select ESM items, 2) the limited ability of current – purely quantitative – ESM design to take subjective meaning making and idiographic context into account and 3) the added value of qualitative assessment for meaningful measurement and integrated validation of ESM designs. Data is presented from a mixed methods ESM pilot study with depressed adolescents. Adolescents' daily experiences were analyzed qualitatively to 1) contextualize person-specific dynamics measured with ESM and 2) examine whether and how ESM can capture relevant domains in adolescents' daily lives. The results of the pilot study will be presented and discussed in light of the broader symposium theme: the challenges and benefits of mixed methodology.

Symposium title

Integrating quantitative and qualitative evidence for developing and validating psychological assessments: Challenges and benefits of mixed methodology

2.5 Parallel sessions 08h30–10h00 Auditorium 2

Symposium Overview

Social Network Methodology

Author(s)

Marijtje van Duijn

University of Groningen, Groningen, Netherlands

Abstract

This symposium presents new developments in the collection and analysis of social network data, along a couple of methodological dimensions. Personal social network data vs. complete social network data; analyses aimed at individual outcomes vs. analyses aimed at generalization at the population level.
All of the contributions combine social networks with existing methods with the aim to improve the measurement or performance of the applied methodology and provide better answers to the research questions.
More specifically, the symposium contains two contributions on including personal social network data in experienced sampling methodology to estimate individual outcomes and provide personal feedback, and two contributions investigating the properties of new applications of existing social network models.

Title

Capturing The Social Life of A Person By Integrating Experience Sampling Methodology And Personal Network Data

Author(s)

Marie Stadel[1], Anna Langener[1], Timon Elmer[2], Gert Stulp[1], Marijtje van Duijn[1], Laura Bringmann[1]

[1]University of Groningen, Groningen, Netherlands; [2]ETH Zürich, Zürich, Switzerland

Abstract

The daily social context of a person is dynamic and contains multiple levels that can be captured with different methodologies. Two methods that seem especially promising when combined are personal social network (PSN) data collection and experience sampling methodology (ESM). While PSN provides data on a person's social relationships and broader social environment, ESM delivers intensive longitudinal data of social interactions (with social network members) in daily life. Despite many potential uses of such multi-layered data, there are few studies to date using the two methods in conjunction.
During this talk, we will illustrate how to combine repeated PSN data collection and ESM to idiographically assess the social relationships and social interactions of a participant. Furthermore, we show a personalized feedback prototype that interactively visualizes the collected data and provides detailed insights into the participant's social context.

Symposium title

Social Network Methodology

Title

Predicting mood based on the social environment measured through social networks combined with experience sampling method and digital phenotyping

Author(s)

Anna Langener, Laura Bringmann, Martien Kas, Gert Stulp

University of Groningen, Groningen, Netherlands

Abstract

Social interactions are essential for mental health. Therefore, researchers increasingly attempt to capture an individual's social environment to predict and explain changes in well-being. Digital phenotyping is an often used technology to assess a person's social behavior through passive sensing and without self-report. Additionally, the experience sampling method (ESM) can capture the subjective perception of specific interactions multiple times per day. Lastly, egocentric networks are often used to measure specific relationship characteristics. Although those different methods capture different aspects of the social environment that are related to well-being, they have rarely been combined in previous research. Combining those methods may be necessary to increase the predictive accuracy of well-being and thus its utility for clinical applications.

In this study, we aim to investigate how accurately we can predict mood based on the social environment as measured through digital phenotyping, ESM, and ego-centric networks. We examine how much each of those three methods adds to the prediction, which allows us to identify the most important measurements for predicting health outcomes.

We use data from a student sample collected over a 28-day period. We train individualized machine learning models and calculate feature importance scores. Overall, we investigate how feasible it is to predict mood, which might be useful for developing just-in-time interventions. Furthermore, identifying which parts of the social environment are most relevant, can help to deliver personalized interventions and to reduce the participant burden.

Symposium title

Social Network Methodology

Title

The Network Scale-Up Method: Validity and reliability

Author(s)

Miranda Lubbers[1], Michał Bojanowski[1,2]

[1]Universitat Autònoma de Barcelona, Barcelona, Spain; [2]Kozminski University, Warsaw, Poland

Abstract

The Network Scale-Up Method (NSUM) was originally designed to estimate the size of hard-to-count populations, but has also been adopted by social network analysts to estimate individuals' acquaintanceship volume (i.e., extended network size) and the social cohesion of their networks across categorical boundaries. Consequently, the NSUM method has been widely used in health sciences and network analysis, and good practices have been developed. Despite its prevalence, as far as we know, the test-retest reliability of NSUM modules has not been estimated yet and more work is needed to establish its internal consistency and validity. In this paper, we discuss the results of a pretest with 50 participants in Spain in 2023, who were interviewed twice over a period of 10-14 days. In both occassions, they responded to an NSUM instrument with several qualitative and quantitative follow-up questions. At the second measurement, a few new items were added to the NSUM instrument. We discuss the measure´s reliability and validity and its implications for use in social network studies and health studies.

Symposium title

Social Network Methodology

Title

Social network meta analysis

Author(s)

Marijtje van Duijn

University of Groningen, Groningen, Netherlands

Abstract

In the past two decades the development of statistical models for the analysis of complete social networks and accompanying software, have led to an increase in the collection of cross-sectional and longitudinal social network data to investigate many relevant research questions. To improve the answers to these questions, we need more knowledge on how to adequately summarize the results of multiple social network analyses. A good and well-known option is to use meta-analysis to obtain an overall estimate of the various parameters in the statistical model applied for the social network analysis.
An important underlying assumption is that the networks are comparable, i.e. that the same statistical network model can be specified for each single network, and will result in reliable parameter estimates. Put differently, it is assumed that the sample of networks are drawn from the same distribution. Because the size of the network and the parameters are not independent, this assumption requires further attention.
In a simulation study based on a sample of classroom networks, the behavior of the estimates of the parameters and goodness of fit statistics of the exponential random graph model (ERGM) are investigated.

Symposium title

Social Network Methodology

2.6 Parallel sessions 08h30–10h00 Auditorium 3

Title

Skill-dependent gender preferences in recruitment

Author(s)

Szymon Czarnik, Marcin Kocór

Jagiellonian University, Krakow, Poland

Abstract

The data for the analysis were collected within the framework of Human Capital Study, 2010-2014 (first edition) and 2017-2022 (second edition). Two of the major components of the study have been surveys of working-age population and companies active in the Polish labour market. The size of the samples (in the first edition, 17,700 persons and 16,000 companies each year) offers a unique and detailed insight into the situation of job-seekers and the recruitment policies of the companies. Employers seeking people for particular job positions were asked about their preferences regarding level of education, age and gender of the prospective employees. We analyze determinants of gender preferences in recruitment with particular focus on the combination of skills required for the advertised positions. These skill-dependent preferences are then set against the skill self-evaluations of men and women seeking employment in order to establish the degree of correspondence between the gendered patterns of demand and supply side in the labour market.

Oral presentations session title:

Applications in HR and Personality

Title

Latent Class Markov Modelling for Studying Dynamic Organisational Configurations: Trajectories of New Venture Competitiveness

Author(s)

Paul Steffens[1], Leo Paas[2], Scott Gordon[1]

[1]University of Adelaide, Adelaide, Australia; [2]University of Auckland, Auckland, New Zealand

Abstract

Interest in configurational approaches have seen a resurgence in entrepreneurship and organisational studies over recent years. Techniques such as QCA and fsQCA commonly used to analyse organizational configurations are static in nature, and temporal extensions are limited. However, many issues of interest to entrepreneurship scholars, such as business model innovation or change, organizational renewal and new venture development are concerned with changes in organizational configurations over time. The authors report a novel application of latent class Markov modelling (LCMM) that facilitates dynamic analysis of organizational configurations. They illustrate the utility of the application by analyzing changes in the configuration of resource-based competitiveness for longitudinal samples of nascent and operational new ventures. We adopted the methods pioneered by the PSED study (Reynolds & Miller, 1992) and subsequently adopted by many panel studies of early-stage ventures. A large-scale phone screening survey of 30,193 randomly selected Australian adults was used to identify a sample of early-stage ventures – nascent ventures (actively working on starting a venture) and operational ventures (less than 3.5 years old). This resulted in 625 nascent ventures and 559 operational ventures that completed Wave 1. Our study utilises three annual waves of data collection, including ventures that are known to have exited in Waves 2 or 3, but excluding ventures that did not complete surveys in Waves 2 or 3.

Oral presentations session title:

Applications in HR and Personality

Title

Do different versions of the same instrument capture the same? An illustration with the two, eight and nine items versions of the Patient Health Questionnaire for assessing major depression in Primary Health Care

Author(s)

Irene Gómez-Gómez[1,2], Isabel Benítez[3,4], Juan Bellón[5,2,6,7], Patricia Moreno-Peral[8,2], Bárbara Oliván-Blázquez[9,10,2], Ana Clavería[11,12,2], Edurne Zabaleta-del-Olmo[13,14,15,2], Joan Llovera[16,17,2], María J. Serrano-Ripoll[16,17,2], Olaya Tamayo-Morales[18], Emma Motrico[1,2]

[1]Department of Psychology, Universidad Loyola Andalucía, Seville, Spain; [2]Prevention and Health Promotion Research Network (redIAPP)/Network for Research on Chronicity, Primary Care, and Health Promotion (RICAPPS), Barcelona, Spain; [3]Department of Methodology of Behavioral Sciences, University of Granada, Granada, Spain; [4]Mind, Brain and Behaviour Research Center (CIMCYC), Granada, Spain; [5]Biomedical Research Institute of Málaga (IBIMA-Bionand platform), Málaga, Spain; [6]El Palo Health Centre, Andalusian Health Service (SAS), Málaga, Spain; [7]Department of Public Health and Psychiatry, University of Málaga, Málaga, Spain; [8]Personality, evaluation and psychological treatment, University of Málaga, Málaga, Spain; [9]Department of Psychology and Sociology, Universidad de Zaragoza, Zaragoza, Spain; [10]Institute for Health Research Aragón (IISA), Zaragoza, Spain; [11]Primary Care Research Unit, Área de Vigo, SERGAS, Vigo, Spain; [12]I-Saúde Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Vigo, Spain; [13]Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain; [14]Atenció Primària Barcelona Ciutat, Gerència Territorial de Barcelona, Institut Català de la Salut, Barcelona, Spain; [15]Nursing department, Faculty of Nursing, Universitat de Girona, Girona, Spain; [16]Primary Care Research Unit of Mallorca, Balearic Islands Health Services, Palma de Mallorca, Spain; [17]Health Research Institute of the Balearic Islands (IdISBa), Palma de Mallorca, Spain; [18]Unidad de Investigación en Atención Primaria de Salamanca (APISAL), Instituto de Investigación Biomédica de Salamanca (IBSAL), Salamanca, Spain

Abstract

In Primary Health Care (PHC) users, the prevalence of major depression is 9.6%. However, PHC professionals only recognize 50% of cases. Easy and quick application instruments are needed for screening major depression in PHC. The Patient Health Questionnaire (PHQ) is the most widely used; however, there is not enough evidence about the functioning of the three available versions when applied to PHC users. The present study aims to assess the psychometric properties of the PHQ-2, PHQ-8 and PHQ-9 for screening major depression in Spanish PHC. A total of 2579 Spanish PHC users from 22 PHC centres participated. Reliability was estimated through McDonald's omega coefficient and its 95% confidence interval. Validity evidence about internal structure, convergence between versions and relations to other variables such as the Composite International Diagnostic Interview (CIDI) used as gold-standard were collected. McDonald's omega coefficient was 0.83 for both the PHQ-8 (95% CI 0.81 to 0.84) and the PHQ-9 (95% CI 0.82 to 0.85). One factor solution was found through an Exploratory Factor Analysis and confirmed by a Confirmatory Factor Analysis. The expected relationships between PHQ-2, PHQ-8 and PHQ-9 and other related measures (anxiety, social support, quality of life and depression) were also confirmed. The optimal cut-off values to detect major depression were 2 for PHQ-2, 7 for PHQ-8 and 8 for PHQ-9. The PHQ

is an adequate instrument for screening major depression in Spanish PHC users. Differences identified between versions will be discussed.

Oral presentations session title:

Applications in HR and Personality

Title

Ads Quality Experiments Using Human Evaluation at Google

Author(s)

Jeremy Miles, Youwei Li

Google, Los Angeles, USA

Abstract

Google tries to show ads that are relevant and useful - thereby improving user, publisher and advertiser satisfaction.To this end, we run experiments to determine the best ads to associate with searches. These automated experiments have potentially interesting parallels and differences with randomized trials in social sciences, which we discuss here.
For a search ads experiment, the unit of analysis is the search query. First we take a weighted sample of queries, and their associated ads. We do not randomize, each query receives experiment and control treatment, so the counterfactual is observable.
Ads are sent to human raters who score the ad from -100 to 100. Each ad is rated several raters and a debiased, aggregated score is calculated using a TensorFlow Probability mixed effects model. Raters are random, ads are fixed effects. Reliability / agreement is assessed using Intraclass Correlation and Krippendorff's Alpha.
Differences between treatment and control are assessed. The Average Treatment Effect on the Treated (ATT) is calculated as the difference, and the Average Treatment Effect (ATE) can be calculated accurately because of the known counterfactual and the sampling scheme. This is converted to a relative effect by comparing the effect size with a population estimate.
Code runs automatically, without human supervision and must provide appropriate estimates, and confidence intervals. We use (bucketed) jackknife or bootstrap procedures to estimate the confidence intervals..
We finish with a brief discussion of the difficulty of designing experiments for the detection of differences in numbers of rare but egregiously bad ads.

Oral presentations session title:

Applications in HR and Personality

2.7 Parallel sessions 08h30–10h00 Auditorium 4

Title

Bayesian auto-regressive dependence latent growth modeling; a novel framework depicted with covid-19 data.

Author(s)

Zachary Roman[1], Holger Brandt[2]

[1]University of Zurich, Zurich, Switzerland; [2]University of Tuebingen, Tuebingen, Germany

Abstract

Spatial dependence occurs when cases in a statistical analysis have a systematic relationship on a variable in space. For example, cultural traits tend to be more similar in nearer regions. Social dependence occurs when cases who are more socially engaged share a systematic relationship on a variable. For example, close friends have a stronger impact on opinion formation as compared to acquaintances. These forms of dependence are different than dependence due to time because of the potential for reciprocal effects. Dependence in time is unidirectional, only earlier measurements/ events can effect later events, in space or social groups nearby regions or socially engaged individuals can effect one another reciprocally. For example, infectious diseases or violent crime can spillover from region A to nearby region B and vice versa. Likewise, opinions of contraceptive use or alcohol abuse of person Acould impact the opinions of a friend, person B, or vice versa. Even further, case A could impact case B in turn caseB can impact case C, extrapolated to more cases this process is referred to as spillover.
One way to accommodate, measure, and test hypothesis related to both forms of spillover are referred to as spatial/social network autoregressive models. In Roman & Brandt (2021) we established the benefits of integrating these autoregressive effects with cross-sectional structural equation models. In this talk I present a Bayesian latent growth curve specification which can model autoregressive dependence with temporal interactions. Simulation results and an empirical application with Covid-19 data will be presented and discussed.

Oral presentations session title:

Latent Variable Models

Title

Factor Score Vector Autoregression: A Two-step Approach to Autoregressive Modeling with Latent Variables

Author(s)

Manuel T. Rein[1], Jeroen K. Vermunt[1], Kim De Roover[2,1], Leonie V.D.E. Vogelsmeier[1]

[1]Tilburg University, Tilburg, Netherlands; [2]KU Leuven, Leuven, Belgium

Abstract

More and more researchers study the trajectories of latent variables across time, for example, to what extent emotions carry over and interact with each other from one moment to the next. When analyzing such dynamic processes, researchers often apply a stepwise procedure in which they first compute sum or factor scores of multiple items and then use them in autoregressive modeling. However, sum scores rely on strict assumptions about the measurement model (indicating how constructs are measured by items) and ignore possible between-person differences therein. Using factor scores requires less assumptions to hold and considers measurement model differences, but neglects the uncertainty in the factor score estimates. In both cases, the resulting autoregressive parameters may be biased. Factor Score Path Analysis addresses these issues but does not readily accommodate longitudinal data. Dynamic Structural Equation Modeling (DSEM) is the state-of-the-art approach for analyzing dynamic processes while including the measurement model. However, it does not offer the stepwise approach researchers may prefer. We propose Factor Score Vector Auto-Regression that extends Factor Score Path Analysis to Vector Autoregressive Modeling. This two-step method provides an alternative to DSEM and can be estimated in lavaan. First, the (possibly partially person-specific) measurement model of each construct is evaluated using factor analysis, and factor scores are computed. Second, the factor scores are regressed on those from the previous observation while taking into account their inherent uncertainty. Through a simulation study, we demonstrate that the method performs well in obtaining correct parameter estimates of a bivariate dynamic process.

Oral presentations session title:

Latent Variable Models

Title

Longitudinal Confirmatory Factor Analysis Model (LCFAM) Misspecification and its impact on the performance of fit indexes and on the Curve-of-Factor Model (CFM)

Author(s)

Elizabeth Valeriano-Lorenzo[1,2], Carmen Ximénez[1], Teodoro del Ser[2]

[1]Autonoma University of Madrid, Madrid, Spain; [2]CIEN Foundation, Queen Sofia Foundation Alzheimer Centre, Madrid, Spain

Abstract

This work presents the results of a study exploring the impact of model misspecification when estimating the measurement model in the context of a Curve-of-Factor Model (CFM), which is a variation of a Latent Growth Curve Model (LGCM) and it is a helpful statistical method that allows the modeling of the change across time. More specifically, we focus on the Longitudinal Confirmatory Factor Analysis Model (LCFAM). In the first step, a measurement model (LCFAM) is tested. Secondly, a structural model is developed to fully assess the extent of the CFM.

We present the results of a Simulation study exploring the impact of Model misspecification in LCFAM on two dependent variables: 1) the performance of the sensitivity of goodness-of-fit indices, and 2) the performance of the CFM measuring the recovery of parameters. The misspecification conditions consisted of 1) autocorrelated errors of the observed variable across time, and 2) indicator-specific variance in the same factor over time. The present work also applies the research recommendations about the use of unbiased SRMR (Maydeu-Olivares, 2017; Ximénez, Maydeu-Olivares, Shi & Revuelta, 2022).

A simulation study was performed to examine the hypothesis that the effect of Model misspecification is more pronounced as the sample size decreased. In addition, the proposed approach is illustrated with empirical data about degenerative disease from a longitudinal cohort.

It must be highlighted that although CoFM is a helpful tool to test theoretical hypotheses about dynamics phenomena, the misspecification of the longitudinal measurement model (LCFAM) influences the results and the conclusions made.

Oral presentations session title:

Latent Variable Models

Title

New Perspectives on Latent Markov Factor Analysis: Embracing Measurement Model Dynamics in Intensive Longitudinal Data

Author(s)

Leonie V.D.E. Vogelsmeier

Tilburg University, Tilburg, Netherlands

Abstract

Research is increasingly moving toward intensive longitudinal methods to look at dynamics in psychological constructs such as affect and well-being. Still, researchers typically assume that the measurement model (MM)—the way items relate to underlying latent constructs—is notchanging across time (i.e., longitudinal measurement invariance holds) or see these changes as nuisance. Not only is this often not realistic, but it completely ignores the valuable insights that can be gained from embracing and teasing apart the dynamics of the MM itself and how it relates to characteristics of the individuals (e.g., personality) and the contexts (e.g., onset of a stressful event). Studying MM dynamics can uncover context-specific changes in, for example, emotional granularity, engagement, and affect experience and ultimately contribute to theory building. The latter is crucial in emerging fields like intensive longitudinal research that are still light on theory. A novel method ideally suited for uncovering these dynamics is Latent Markov factor analysis (LMFA; Vogelsmeier et al., 2019), which combines a discrete- or continuous-time latent Markov model (that clusters observations into separate states, according to state-specific MMs) with mixture factor analysis (that evaluates which MM applies for each state). In this presentation, I will describe how LMFA reveals MM differences across individuals and time as well as possible reasons for these dynamics, and illustrate LMFA with empirical applications. I will also introduce the new user-friendly software package "lmfa" that allows researchers easily embrace MM dynamics in their own intensive longitudinal data.

Oral presentations session title:

Latent Variable Models

2.8 Parallel sessions 08h30–10h00 Lecture room 1.2

Title

A new sample size planning approach for (V)AR(1) models: Predictive Accuracy Analysis

Author(s)

Jordan Revol, Ginette Lafit, Eva Ceulemans

KU Leuven, Leuven, Belgium

Abstract

Researchers often use intensive longitudinal designs in combination with VAR(1) models to capture processes that evolve dynamically in time. In this context, sample size planning (i.e., number of measurement occasions needed) often uses power as criterion. One drawback is that power-based sample size recommendations will depend on the inspected model parameter at hand and will not hold for the model as a whole. Moreover, power analysis takes an explanatory stance while the predictive stance, that focuses on the performance of the full model for predicting unseen data, is increasingly used as well. We therefore suggest to consider predictive accuracy as a sample size planning criterion. Focusing on VAR(1) in a N=1 context, we propose a novel predictive accuracy metric and a new simulation-based method, called predictive accuracy analysis (PAA), to assess how many measurement occasions are required in order to optimize predictive accuracy. Specifically, we introduce a new predictive accuracy metric which is based on the VAR(1) assumptions and on the expected 'true' parameter values, after standardizing them by computing the Mahalanobis distance between them and the true error distribution. This distance measure allows to account for the innovation covariances of the processes. Hence, our predictive accuracy analysis computes the sample size required so that the proportion of simulated replicates (e.g., .8) in which the inspected proportion of multivariate prediction errors of acceptable size is high enough. Finally, we showcase how the different VAR(1) model parameters impact sample size planning recommendations.

Oral presentations session title:

Intensive Longitudinal Data

Title

One does not simply correct for serial dependence

Author(s)

Sigert Ariens, Janne Adolf, Eva Ceulemans

KU Leuven, Leuven, Belgium

Abstract

Serial dependence is present in most time series data sets collected in psychological research. This paper investigates the implications of various approaches for handling such serial dependence, when one is interested in the linear effect of a time-varying covariate on the time-varying criterion. Specifically, the serial dependence is either neglected, corrected for by specifying autocorrelated residuals, or modeled by including a lagged version of the criterion as an additional predictor. Using both empirical and simulated data, we showcase that the obtained results depend considerably on which approach is selected. We discuss how these differences can be explained by understanding the restrictions imposed under the various approaches. Based on the insight that all three approaches are restricted versions of an autoregressive distributed lag model, we demonstrate that accessible statistical tools, such as information criteria and likelihood-ratio tests can be used to justify a chosen approach empirically.

Oral presentations session title:

Intensive Longitudinal Data

Title

Specifying models for between-individual differences in longitudinal data

Author(s)

Anja Ernst, Casper Albers, Marieke Timmerman

University of Groningen, Groningen, Netherlands

Abstract

Across different fields of research the similarities and differences between various longitudinal models are not always eminently clear due to differences in data structure, application area, and terminology. I will present a comprehensive model framework that allows simple comparisons between longitudinal models, to ease their empirical application and interpretation. At the within-individual level this model framework accounts for various attributes of longitudinal data, such as growth and decline, cyclical trends, and the dynamic interplay between variables over time. At the between-individual level the framework contains continuous and categorical latent variables to account for between-individual differences. This framework encompasses several well-known longitudinal models, including multilevel regression models, growth curve models, growth mixture models, vector-autoregressive models, and multilevel vector-autoregressive models.
In my talk I will illustrate key characteristics of the framework using famous longitudinal models as concrete examples. Being familiar with these key characteristics can aid interpretation and comparisons across different longitudinal models. I will make recommendations for selecting and specifying longitudinal models for researchers who aim to account for between-individual differences.

Oral presentations session title:

Intensive Longitudinal Data

Title

Improved estimation of autoregressive models through contextual impulses and robust modeling

Author(s)

Janne Adolf, Eva Ceulemans

KU Leuven, Leuven, Belgium

Abstract

The aim of the dynamic paradigm of affect research is to characterize daily life affective processes by means of intense longitudinal data and dynamic, typically autoregressive-type models. The contextual conditions accompanying and potentially influencing affective processes obviously form an important part of the picture. Especially distinct contextual events – ranging from salient daily events to major life events – are regularly assessed and their effects modelled. Such an explicit approach to studying context can however reach its limits, if relevant events are hard to define, measure or model. In that case one finds oneself in a situation where contextual events play out as hidden contaminators possibly obscuring the affective process of interest. Interestingly, such contamination can also have beneficial effects in that it can trigger autoregressive dynamics and leverage their estimation. In this talk we take a closer look at this phenomenon. We also demonstrate that robust autoregressive models deal especially well with contextual contamination as they not only capitalize on positive leverage effects but also mitigate the negative obscuring effects contextual events might have.

Oral presentations session title:

Intensive Longitudinal Data

2.9 Parallel sessions 08h30–10h00 Lecture room 1.3

Title

Conceptualization of Myths about Cyber-Sexual Violence Against Women: A Thematic Analysis of Social Reactions to Reports on Twitter

Author(s)

Rocío Vizcaíno-Cuenca[1], Mónica Romero-Sánchez[2], Hugo Carretero-Dios[1]

[1]Department of Behavioural Science Methodology, University of Granada, Granada, Spain; [2]Department of Social Psychology, University of Granada, Granada, Spain

Abstract

Introduction: Sexual violence experienced by women in online spaces represents an understudied phenomenon. Despite the undoubted benefits of social networks, negative consequences have also arisen from the online interactions and its characteristics (i.e., anonymity, wide audience, etc.). Among them, this study analyses "cyber-sexual violence" against women, as well as the sexist online culture that justifies such violence. Specifically, we are interested in gaining an in-depth understanding of the social perception of this type of violence in order to ultimately isolate the content areas that would serve to define the myths about cyber-sexual violence against women. To this end, and following previous research that has highlighted the relevance of information exchanged on social networks (i.e., Twitter), this study conducts a qualitative analysis of social reactions to reports on Twitter.
Method: First, 18 reports of cyber-sexual violence against women published on Twitter were selected using the rtweet data package implemented in the statistical software R. Second, from the reports, a total of 4,048 reactions were extracted using the Octoparse software. Finally, from these reactions, a thematic analysis has been performed following the proposal of Clarke and Braun (2018).
Results: The results of the analysis show that there are attitudes that justify cyber-sexual violence against women through: 1) minimizing, 2) victim blaming, 3) exonerating the perpetrator responsibility and 4) socio-cultural factors.
Conclusion: This study enables a deeper understanding of myths about cyber-sexual violence against women, and it has constituted the first phase in the development of an instrument.

Oral presentations session title:

Research Design and Qualitative Methodology

Title

A methodological proposal to study moral decision-making when facing moral dilemmas in a comprehensive and systematic way.

Author(s)

Belén Carrascal-Caputto[1,2], Noelia Aguilera-Jiménez[3], Araceli Méndez-Andrade[4], José Luis Padilla[1,2], Pilar Aguilar[4], Isabel Benítez[1,2]

[1]University of Granada, Granada, Spain; [2]Brain and Behaviour Research Center (CIMCYC), Granada, Spain; [3]University of Huelva, Huelva, Spain; [4]Universidad Loyola Andalucía, Sevilla, Spain

Abstract

In the last years, the concerns about the accuracy of the inferences related to human morality have increased, partly because of the heterogeneity of the designs used to address the study of that phenomenon. The variety of variables manipulated in the experiments, the lack of replicability and the robustness of the construct definition are some of the limitations highlighted in the literature. In addition, recent studies warn about the lack of convergence between the psychological factors and processes developed when making decisions in the laboratory through sacrificial dilemmas and when facing dilemmas in real life. The objective of the present study is to explore and identify the most effective procedural and methodological factors to study decision-making in the face of moral dilemmas. To achieve this goal, two studies were conducted. First, a systematic review of mixed studies focused on collecting and synthesizing the procedures used in empirical studies on moral decision-making. Subsequently, a qualitative study where real moral dilemmas were compiled to identify the factors involved in making moral decisions in real life. Results from both sources were integrated to reach a list of elements to be considered when assessing moral decision-making. A proposal to comprehensively approach to the study of moral dilemmas will be presented and discussed, as well as implications and contributions related to standardized the study of moral decision-making.

Oral presentations session title:

Research Design and Qualitative Methodology

Title

Advancing research through a non-positivist quantitative methodology embedded in four levels of research context

Author(s)

Regine Haardoerfer, Melvin Livingston

Emory University, Atlanta, USA

Abstract

To advance the decolonization of research, we have adapted non-positivist methodologies already common in qualitative research to advance quantitative practice. This methodology builds upon the seminal work Indigenous Statistics: A Quantitative Methodology by Walter and Andersen (2013). It includes the researcher's position as composed of their philosophy (ontology, axiology, and epistemology) and social status which is (consciously or subconsciously) the foundation for theoretical frameworks and choice in research methods. We have embedded the researcher and their research in four levels of context as proposed by Scheurich (1997): individual, institutional, societal, and civilizational which all need to be scrutinized in order to address the multitudes of oppression filtering into our research (e.g., racism, colonialism, sexism, ableism, etc.). To move from describing the methodology into action, we propose two key elements that we see as opportunities. First, we should endeavor to diversify our research teams. This can be done through engaging in critical reflection on the positionality of both the team as a whole and team members as individuals. This reflection should be ongoing throughout the research process to better facilitate minimizing our collective bias in quantitative decision making; which in turn will yield better and more actionable research. Second, the use of strong participatory methods, such as community-based participatory research (CBPR), which center communities' knowledges, resources, and strengths. By centering community strengths, quantitative research can better focus on pragmatic and culturally meaningful solutions.

Oral presentations session title:

Research Design and Qualitative Methodology

Title

A qualitative evaluation of an (quasi)-experiment: Studying the effects of empathy-inducing probes on distancing during Covid to derive methodological guidelines.

Author(s)

Hidde Leplaa[1], Karlijn Soppe[2], Jari Tonjes[1], Mariska Bouterse[1], Irene Klugkist[1]

[1]Utrecht University, Utrecht, Netherlands; [2]Ghent University, Gent, Belgium

Abstract

Experiments and quasi-experiments are almost invariably evaluated with quantitative methods. We argue that there can be added value of using qualitative methods to evaluate an experiment. In the context of a larger study, we explored and explained the methodological steps of the qualitative evaluation of a potential intervention effect. The study was conducted during the Covid-crisis, and investigated the effect of Empathy-inducing probes on the distance kept between people, measured by photographs taken at intervals. Within this context there was room to conduct our qualitative study. We focused on both actual behavior (referred to as strategies in our study) and intentions (motivations). We collected data using observations and two types of interviews. The analysis of the qualitative data was done following the constructivist approach to the Grounded theory. Our study provided guidelines for each step of a qualitative evaluation of an experiment: 1) the formulation of the research goals, 2) data collection, 3) data analysis, 4) interpretation of the intervention effect, and 5) ensuring the rigor of the research. We conclude that, by adding a qualitative analysis method to an (quasi-)experiment the ecological validity of a study can be enhanced, by acquiring a more holistic understanding of the phenomenon of interest.

Oral presentations session title:

Research Design and Qualitative Methodology

2.10 Parallel sessions 1h30–13h00 Auditorium 1

Symposium Overview

Methodological Advances in Meta-analysis

Author(s)

Julio Sánchez-Meca[1], Juan Botella[2]

[1]University of Murcia, Murcia, Spain; [2]Autonomous University of Madrid, Madrid, Spain

Abstract

Meta-analysis is a research methodology that is continuously growing to adapting to many different empirical studies and synthesis proposals. This symposium presents four recent developments in the methodology of meta-analysis representing researchers from the University of Leuven (Wim van den Noortgate), Maastricht University (Wolfgang Viechtbauer), National University of Distance Education (UNED) in Madrid (Belén Fernández-Castilla), Complutense University of Madrid (Raimundo Aguayo), Autonomous University of Madrid (Juan Botella, Manuel Suero, Juan I. Durán), Distance University of Madrid (UDIMA; Desirée Blázquez-Rincón), and University of Murcia (Julio Sánchez-Meca, José A. López-López, Alejandro Veas-Iniesta, María Rubio-Aparicio, José Antonio López-Pina). A first talk is presented by José A. López-López on how to manage heterogeneity in meta-analysis by means of a new kind of statistical models named scale-location meta-regression models. In the second talk, Juan Botella presents a new formulation of the random-effects model when the standardized mean difference is the effect size index, that better solve the problems of the standard random-effects model. In the third talk, Alejandro Veas-Iniesta presents the Meta-Analytic Structural Equation Modeling (MASEM) applied to reliability generalization meta-analysis and illustrates it with an application. Finally, Belén Fernández-Castilla presents a systematic review of the studies that have applied network meta-analysis in Psychology and Educational Sciences, with the purpose of identifying potential caveats in the reporting practices.

Title

Implementation of Location-Scale Models in Meta-Analysis: A Simulation Study

Author(s)

Desirée Blázquez-Rincón[1], Wolfgang Viechtbauer[2], José Antonio López-López[3]

[1]Distance University of Madrid, Madrid, Spain; [2]Maastricht University, Maastricht, Netherlands; [3]University of Murcia, Murcia, Spain

Abstract

Purpose: Location-scale models for meta-analysis allow simultaneous testing of moderators of the mean (location) and variance (scale) of the distribution of true effects. Different methods recently implemented in the metafor R package for fitting meta-analytic location-scale models are compared. Method/Design: Monte Carlo simulation aimed to compare different estimation methods (maximum or restricted maximum likelihood estimation), significance tests (Wald-type, likelihood-ratio, or permutation tests), and confidence intervals for scale coefficients (Wald-type or profile-likelihood intervals). We examined the impact of the number of studies, study size, amount of heterogeneity, and type of moderator on the rejection rates of the tests and on the coverage and width of the intervals. Results: The permutation test achieved Type I error rates closer to the nominal 5% level than the likelihood-ratio and Wald-type tests, with the likelihood-ratio showing the highest statistical power rates out of all three tests. Compared to the Wald-type method, the profile-likelihood method yielded narrower intervals that were also closer to the nominal 95% level in most scenarios. Although the estimation method made little difference for most conditions, results using restricted maximum likelihood performed closer to the nominal level than maximum likelihood in scenarios with small to moderate number of studies and large heterogeneity. Rejection rates and coverage probabilities were slightly higher when a qualitative moderator was examined rather than a quantitative one. Conclusions: Location-scale models are a powerful tool for meta-analysis and can help researchers address questions that have not yet been explored.
Funding. Grant PID2019-104033GA-I00 funded by MCIN/AEI/10.13039/ 50110 00110 33

Symposium title

Methodological Advances in Meta-analysis

Title

Reformulating the meta-analytical random effects model of the standardized mean difference as a mixture model

Author(s)

Manuel Suero, Juan Botella, Juan I. Durán

Autonomous University of Madrid, Madrid, Spain

Abstract

In meta-analysis the effect size (ES) values are usually modeled through a "classical" formulation of the random effects model. When applied to the standardized mean difference, that formulation has a number of weaknesses, linked both to the weighting scheme and the assumptions made. Among them are that the sampling variances are known, that true effects are normally distributed, and a specific interpretation of the marginal variances of the estimates. Although most weaknesses are "negligible", all together conform a weak and not rigorous version of the random effect model. In this communication we discuss an alternative formulation, within the framework of the mixture models. We assess that reformulation, implying: (a) a weighting scheme for the mean effect in which the estimates and the weights do not correlate, (b) an estimate of the specific variance that do not assume arbitrary distributions, and (c) a better specification of the marginal distributions. The formulas derived from that reformulation allow better estimates of the hyper-parameters in conditions where other estimates show larger bias.

Symposium title

Methodological Advances in Meta-analysis

Title

Meta-analysis of Structural Equation Modelling (MASEM): Principles and applications for improving reliability generalization processes in tests and scales

Author(s)

Alejandro Veas[1], José Antonio López-López[1], María Rubio-Aparicio[1], Raimundo Aguayo[2], José Antonio López-Pina[1], Julio Sánchez-Meca[1]

[1]University of Murcia, Murcia, Spain; [2]Complutense University of Madrid, Madrid, Spain

Abstract

In recent years, evidence has highlighted certain limitations of traditional methods used for meta-analyses of reliability generalization, including the lack of equivalence between total and subscale reliability indices, or the violation of the principle of error independence. In this regard, multivariate statistical methods have been developed for a more efficient estimation of measurement instruments, such as meta-analysis of structural equation modelling (MASEM). The advantages of MASEM include the ability to combine correlation matrices from primary studies, and the estimation of factor models, among others. This communication shows the usefulness of this technique through its application to the Emotional Quotient Inventory Youth Version (EQ-i:YV), which is one the most internationally used instrument of emotional intelligence in children and adolescents. A MASEM approach will yield more robust reliability measures using Omega values, which are more appropriate for multidimensional measures than Cronbach's alphas. As a result, we expect to shed light on the psychometric properties of the EQ-i:YV and help drive theoretical development in this field area.

Symposium title

Methodological Advances in Meta-analysis

Title

Network meta-analysis in Psychology and Educational Sciences: A systematic review of their quality and characteristics

Author(s)

Belén Fernández-Castilla[1], Wim van den Noortgate[2]

[1]National University of Distance Education, Madrid, Spain; [2]Catholic University of Louvain, Louvain, Belgium

Abstract

Network meta-analysis (NMA) is a powerful statistical method for synthesizing evidence from multiple studies that investigate the effectiveness of different interventions. While NMA has been extensively utilized in the field of medicine, its application in psychology and educational sciences is relatively infrequent. Systematic reviews that describe published NMAs are primarily focused on the medical field, and the parameters used in simulation studies to test the performance of NMA models are often drawn from these reviews. Considering this, the present study aims to describe the characteristics of NMAs published in the fields of psychology and educational sciences, identify common practices and potential shortcomings, and compare these findings to those of NMAs in the medical field. Additionally, we evaluated the quality of reporting of the retrieved NMAs using the extension of the PRISMA statement to NMAs. A total of 42 NMAs were retrieved. Our analysis revealed that NMAs published in the field of psychology and educational sciences tend to include larger number of studies with smaller sample sizes and more intervention groups. Also, inconsistent effects are observed more frequently, and the preferred effect size is the standardized mean difference (rather than the odds ratio in medicine). While the quality of NMAs in psychology is generally satisfactory, some aspects require improvement, such as the description of the treatment network and geometry, the study of bias across studies, and reporting individual study results.

Symposium title

Methodological Advances in Meta-analysis

2.11 Parallel sessions 1h30–13h00 Auditorium 2

Symposium Overview

Using Structural Equation Models to Analyze Round-Robin Data from Social Networks

Author(s)

Terrence Jorgensen

University of Amsterdam, Amsterdam, Netherlands

Abstract

Study social phenomena from an interpersonal perspective is enabled by round-robin designs, in which each member of a group provides data about every other member—e.g., each student in a classroom indicates how much they like each other student, or each nuclear-family member indicates how secure their relationship is with each other family member. This complex pattern of interdependence among dyadic observations (Y_{ij}: a variable Y measured about person i responding to or interacting with person j) has a social-network structure, which requires sophisticated analytical models to account for interdependence. The linear social relations model (SRM: $Y_{ij} = \mu + P_i + T_j + R_{ij}$) decomposes such interpersonal perceptions into random effects associated with perceivers (P_i), their targets (T_j), and relationship-specific nuances captured by dyad-level residuals (R_{ij}). Sampling several round-robin groups (e.g., families, classrooms) also enables modeling of group-level variance via a random intercept μ_g rather than constant mean μ. A multivariate SRM can estimate correlations among multiple round-robin variables, but fitting theoretical models to explain those relationships requires a larger modeling framework, such as structural equation modeling (SEM). This symposium highlights how to analyze SRM data with SEM via open-source software. The first pair of presentations focus on family-SRM data, where family members can have differential influence within their network. The second pair of presentations compare and evaluate 1- and 2-stage maximum-likelihood estimation methods for analyzing multivariate-SRM data via the social-relations SEM (SR-SEM). All presentations and software examples are provided on the Open Science Framework (OSF): https://osf.io/ahuq6

Title

Toward a general multivariate framework for social network data: An overview of estimation methods for structural social-relations models

Author(s)

Terrence D. Jorgensen

University of Amsterdam, Amsterdam, Netherlands

Abstract

Models of social-network data must account for interdependencies among dyadic observations (Y_{ij}) within a round-robin group (i.e., each group member i responds about or interacts with each other member j). The social relations model (SRM) is a linear decomposition of (approximately) continuous variables into person-level random effects—perceivers (P_i) and their targets (T_j)—and dyad-level relationship effects (R_{ij}). Univariate-SRM analyses investigate relative contributions of each effect's variance component, as well as correlations among person-level effects (generalized reciprocity) and dyad-level effects (dyadic reciprocity). Univariate SRM has been extended to allow predictors of person- and dyad-level effects, and multivariate SRM enables estimating correlations between (person- and dyad-level components of) multiple round-robin variables. Various ad-hoc methods have been used to parsimoniously explain such relationships via regression or structural equation models (SEM). We review two-stage estimation procedures to estimate person- and dyad-level effects, which are then treated as data in a subsequent SEM. The recently developed social-relations SEM (SR-SEM) uses single-stage maximum-likelihood estimation (MLE) to avoid bias introduced by treating estimates as data, but it lacks the flexibility of two-stage approaches—e.g., to model only person- or dyad-level effects. We propose a new two-stage MLE technique using estimated summary statistics as data, which is flexible without sacrificing validity of inferential statistics (or requiring the computational burden of other two-stage solutions) and makes it possible to overcome other remaining limitations (e.g., assuming multivariate normality). We discuss the computation details and implementation in the R package `lavaan.srm`.

Symposium title

Using Structural Equation Models to Analyze Round-Robin Data from Social Networks

Title

Comparing one- to two-stage maximum likelihood estimation for structural equation models of social network data

Author(s)

Aditi Manoj Bhangale, Terrence D. Jorgensen

University of Amsterdam, Amsterdam, Netherlands

Abstract

The social relations model (SRM) is applied to examine dyadic data within social networks. Multivariate SRMs have been modelled using linear mixed models, which are insufficient to estimate structural equation models (SEMs) of complex theories. The social relations SEM (SR-SEM) combines the SRM and SEM, allowing researchers to fit more complex models and test several measurement-related and structural hypotheses about associations among SRM components. The current 'gold standard' for estimating SR-SEMs is a single-stage maximum likelihood estimation (MLE) algorithm implemented in the R package srm. We propose a novel two-stage MLE technique that would enable overcoming some existing limitations (e.g., computational burdens, assuming multivariate normality). Stage-1 of the two-stage estimator is Markov chain Monte Carlo estimation of unrestricted summary statistics of SRM effects. Stage-2 is MLE of constrained SEMs using the Stage-1 summary statistics of SRM effects as input data, incorporating uncertainty about the Stage-1 estimates to adjust Stage-2 SEs and test statistics. Benefits of the two-stage approach include flexibility, computational speed, features available in standard SEM software, analysing (incomplete) discrete or continuous nonnormal data, and fitting complex models that the implemented single-stage MLE has difficulty estimating. We perform a simulation to compare the accuracy and efficiency of single- and two-stage MLE for the ideal scenario: normally distributed complete data, and assess the relative bias, relative efficiency, root mean-square error, and coverage rates of the final parameter estimates of both approaches. We also compare the likelihood ratio test and Brown's residual-based method for testing SR-SEM model fit.

Symposium title

Using Structural Equation Models to Analyze Round-Robin Data from Social Networks

Title

Introducing a more Flexible Social Relations Model for family data

Author(s)

Leila Van Imschoot[1], Lara Stas[1,2], Justine Loncke[3], Ann Buysse[1], Tom Loeys[1]

[1]Ghent University, Ghent, Belgium; [2]Vrije Universiteit Brussel, Brussels, Belgium; [3]Independent Researcher, Ghent, Belgium

Abstract

The family SRM (fSRM) decomposes round-robin measurements into individual, dyadic, and family-level components, while considering the different roles in a family. In its original specification within the SEM framework, all factor loadings between the dyadic measurements and SRM components are fixed to one. This implies that all family members' perceptions are equally important in shaping the family effect, e.g., their shared family culture or climate. We argue that this equal-weights assumption might be too stringent. To this end, we introduce the Flexible fSRM. With the Flexible fSRM, the factor loadings of the family effect are freely estimated. Through simulations, we investigate how the Flexible and original fSRM perform under a violated as well as fulfilled equal-weights assumption. The Flexible fSRM results in good fitting models and unbiased estimators in both conditions. On the other hand, the traditional fSRM yields bad model fits and biased estimators when the equal-weights assumption is not satisfied. Moreover, by reanalyzing data from published fSRM studies we examine how tenable the equal-weights assumption is in real family research settings. Results are discussed in light of the commonly reported absence of substantial family-level variance. Taken together, the model can adequately capture family dynamics and enables researchers to elucidate the association between family members' perceptions and their shared family culture.

Symposium title

Using Structural Equation Models to Analyze Round-Robin Data from Social Networks

Title

How to analyze round-robin family data: the social relations model with roles

Author(s)

Lara Stas[1,2], Felix Schönbrodt[3], Leila Van Imschoot[1], Ann Buysse[1], Tom Loeys[1]

[1]Ghent University, Ghent, Belgium; [2]Vrije Universiteit Brussel, Brussels, Belgium; [3]Ludwig-Maximilians-Universiteit München, München, Germany

Abstract

This presentation introduces the Social Relations Model (SRM) for groups with distinguishable dyad members, such as families where each member has a unique role (e.g., mother, father, oldest and youngest child).
The round robin design, in which every family member rates every other member on the same items, is used to obtain observed dyadic scores. These scores can be decomposed into individual, dyadic and family components using the SRM. The SRM parameters can be estimated using a confirmatory factor analysis, but its statistical complexity may pose a challenge for family researchers. To address this, we developed a user-friendly R-package called fSRM which performs almost automatically those rather complex SRM analyses. With fSRM, one line of R-code suffices to fit the SRM.
With fSRM, researchers can analyze both simple and more complex social relations models with just one line of R-code. The package uses confirmatory factor analysis, based on the R-package lavaan for structural equation modeling (Rosseel, 2012). The fSRM-output provides easy-to-interpret summary tables of SRM variances, variance decompositions, individual and dyadic reciprocities. In addition, SRM means, which can provide valuable information but are seldom reported, can be obtained and easily compared between roles. The package is suitable for both single and multigroup studies, and contains multiple additional options (e.g., estimating intragenerational similarities).
Overall, the fSRM package enables family researchers to overcome the statistical complexity of the SRM and get the most out of their data.

Symposium title

Using Structural Equation Models to Analyze Round-Robin Data from Social Networks

2.12 Parallel sessions 1h30–13h00 Auditorium 3

Title

Measurement Question 1: What is it that you want to measure?

Author(s)

Jan De Houwer

Ghent University, Ghent, Belgium

Abstract

Behavioral sciences examine how the states of systems (e.g., an organism, part of an organism, a group of organisms) are a function of stimuli in their environment. In behavioral sciences, to-be-measured constructs can thus be situated at five levels: (1) states of a system, (2) behavior (i.e., the impact of stimuli on the state of a system), (3) changes in behavior (i.e., changes in the way a stimulus changes the state of a system), (4) behavioral effects (i.e., changes in behavior that are due to stimuli or regularities in the presence of stimuli), and (5) entities that mediate behavior or behavioral effects. The higher the level at which a construct is situated, the more difficult it becomes to measure the construct because (1) valid measurement at higher levels requires valid measurement at lower levels (e.g., measuring changes in behavior also requires multiple measurements of behavior, which in turn requires multiple measurements of states) and (2) constructs at higher levels often involve unobservable entities (e.g., causality, information processing). Psychologists typically aim to measure Level 5 constructs (i.e., informational entities such as mental representations that are assumed to mediate behavior or behavioral effects). Because of the challenges of valid measurement at this level, psychologists are well-advised to re-examine whether measurement at lower levels would suffice to achieve their scientific aims.

Oral presentations session title:

Measurement

Title

SIREN: A Hybrid FA-CFA Procedure to Reduce Acquiescence to Insignificance

Author(s)

Ana Hernandez-Dorado, David Navarro-Gonzalez, Fabia Morales-Vives, Pere Joan Ferrando

Universitat Rovira i Virgili, Tarragona, Spain

Abstract

In the structural and dimensional analyses of test items, Acquiescence (AR) is a potentially distorting variable that might affect the analyses outcomes at multiple levels. It is for this reason that it is necessary to use appropriate control methods that avoid or at least minimize these distortions. Here we present a factor-analytic procedure that makes preliminary corrections or data adjustments before fitting the proposed 'content' solutions. The purpose of this communication is to discuss this new tool, named SIREN.
SIREN is a multi-stage hybrid procedure, which combines the CFA and EFA features, and is: (a) designed to fit multiple content solutions and (b) based on scales that will not generally be balanced. The procedure considers the strengths and weaknesses of the two main AR control procedures: fully confirmatory method (Billiet and McClendon, 2000) and exploratory o semi-confirmatory method (Lorenzo-Seva and Ferrando, 2009), and tries to combine the best of both.
A simulation study focused on the recovery of the true loadings and the goodness-of-fit results has been carried out; the goal was to evaluate its functioning under problematic situations for both CFA and EFA solutions.
The results indicate that the procedure is robust and a good alternative to consider. The difference between the estimated loadings and the true loadings is small and the goodness-of-fit methods are stable throughout the replications. No significant differences in loading recovery or model-data fit were found across the different situations that were tested.

Oral presentations session title:

Measurement

Title

Uncovering latent classes to undertand DIF by LCA in 2018 PISA Students Questionnaire.

Author(s)

M.Carmen Navarro-González[1], José-Luis Padilla[1], Luis-Manuel Lozano[1], Álvaro Postigo[1,2]

[1]University of Granada, Granada, Spain; [2]University of Oviedo, Oviedo, Spain

Abstract

In Spain, 24.4% of students reported having detected at least one case of bullying in their class (Fundación ANAR, 2022). The PISA 2018 found that in the majority of countries, boys reported having suffered more bullying than girls, but also boys reported having a stronger sense of belonging to school than girls (OECD, 2018). Those apparently "contradictory" findings could be explored by applying a mixed-methods approach by conducting Differential Item Functioning (DIF) analyses and qualitative methods, to discard the possibility of these findings being a measurement "artifact". It becomes necessary to delve into the response processes to scale items. The study aim is to identify groups of adolescents with different response patterns in the "Experiences of Bullying Scale" from the PISA 2018 Student Questionnaire. Responses of 11599 Spanish students were analyzed by Latent Class Analysis (LCA). We opted for a 4-classes solution based on fit statistics and conceptual criteria. The classes are: a) Students that had not received any type of bullying (72.62%); b) Students that had received all types of bullying (6.28%); c) Students that had only received relational bullying (10,34%); and d) Students that seemed to had received bullying from their own friends (10,76%). The influence of auxiliary sociodemographic covariates in class membership was examined. All analyses were carried out with Mplus (Muthén & Muthén, 1998-2017). In addition, we discuss implications of results for validity of comparative group interpretations.

Oral presentations session title:

Measurement

Title

Detecting and correcting faking in forced-choice personality assessments

Author(s)

Anna Brown

University of Kent, Canterbury, United Kingdom

Abstract

Personality and similar attributes are almost exclusively assessed using self-report questionnaires, which are open to respondents creating the best impression (aka 'faking good'). To counteract faking, the use of 'forced-choice' questionnaires has been popular since appropriate item response models became available to scale them, for example Thurstonian IRT models (Brown & Maydeu-Olivares, 2011, 2018). Forced-choice questionnaires are particularly effective at preventing faking when all choice alternatives appear equally desirable (Cao & Drasgow, 2019). However, to evaluate the effectiveness of matching alternatives on desirability (and potentially correcting for the lack thereof), we need methodology for detecting and measuring faking in forced-choice questionnaires.
OBJECTIVES
This research aims to describe a response model for faking in forced-choice questionnaires, akin to the recent model of 'intermittent' faking for rating scales (Brown & Böckenholt, (2022).
METHOD
I use two-level factor mixture models, with person responses to each set of alternatives at the within level, and personal attributes at the between level. This approach allows modelling faking as person-by-item interactions, taking to account both item properties (such as 'valence' and 'evaluative strength'; Holtgraves, 2004) and personal styles and cognitive abilities important for faking. Every item response can fall into either 'real' or 'ideal' latent class at the within level, measuring the grade of membership in either at the between level.
The approach is illustrated with an operational recruitment study of software engineers (N=2,000), in which stakes are manipulated. It is shown that latencies and other response behaviours can be analyzed in the same analysis framework.

Oral presentations session title:

Measurement

2.13 Parallel sessions 1h30–13h00 Auditorium 4

Title

A Two-Stage Path Analysis Approach to Model Interaction Effects for Congeneric Measures

Author(s)

Gengrui Zhang, Hok Chio (Mark) Lai

University of Southern California, Los Angeles, USA

Abstract

We investigated the performance of a two-stage path analysis method (2S-PA) of modeling and testing interaction effects with congeneric measures in which items under a unidimensional factor model have different factor loadings and error variances. By separating the estimation of factor scores and path coefficients, the 2S-PA allows for different estimation methods of factor scores and path coefficients, and largely avoids model convergence issues by simplifying path models. The 2S-PA can account for interaction effect by multiplying the factor scores of corresponding latent variables and constraining the error variances using standard errors.
A closely related method widely used in current research for estimating interaction effect is reliability-adjusted product indicator (RAPI) method, in which each indicator of the latent focal predictor (e.g., X) is multiplied with an indicator of the latent moderator (e.g., Z) and these products are used as indicators of a new latent variable representing the interaction between the two latent predictors (e.g., XZ). One key limitation of RAPI is that it assumes the observed indicators are continuous and normally distributed with constant measurement error variance, while the 2S-PA method can adjust nonconstant measurement error variances and apply to categorical variables. In the simulation study, we intend to compare the performance of reliability-adjusted product indicator (RAPI) methods to the 2S-PA method, on testing interaction effects, in terms of model convergence and a few model evaluation indices (e.g., standardized bias).

Oral presentations session title:

Structural Equation Modeling (SEM)

Title

Measuring and Comparing the Fitting Propensity of Factor Models

Author(s)

Martina Bader, Morten Moshagen

Ulm University, Ulm, Germany

Abstract

Model selection is an omnipresent issue in structural equation modeling (SEM). When choosing among multiple competing models, a trade-off is often sought between goodness-of-fit and model parsimony. Whereas traditional measures of model fit in SEM quantify parsimony as the number of free parameters estimated by a model, the ability of a model to account for diverse data patterns—known as fitting propensity—also depends on its functional form. This talk describes how the fitting propensity of structural equation models can be investigated by assessing their fit to correlation matrices drawn uniformly from the data space. In addition, the findings of a study comparing the fitting propensity of commonly employed factor models (i.e., exploratory and confirmatory factor analysis models positing a different number of latent factors or a different structure) are reported. It was also investigated to which extent typically used fit indices (CFI, SRMR, RMSEA, and TLI) and information criteria (AIC and BIC) account for differences in fitting propensity. Whereas the compared models exhibited varying degrees of fitting propensity, these differences were mostly due to differences in the number of parameters estimated. Accounting for the number of free parameters via parsimony-adjusted fit indices allowed for an adequate control of differences in the fitting propensity of the compared models.

Oral presentations session title:

Structural Equation Modeling (SEM)

Title

The effect of high subset homogeneity on structural investigations by confirmatory factor analysis

Author(s)

Karl Schweizer, Andreas Gold, Dorothea Krampen

Goethe University Frankfurt, Frankfurt, Germany

Abstract

An investigation of how high subset homogeneity (HSH) influences the outcome of factor analysis of manifest variables expected to represent the same construct is presented. High subset homogeneity denotes a configuration of variables that is inhomogeneous in the following way: some of the variables show a degree of homogeneity that is higher than average whereas the degree of homogeneity among the remaining variables is lower than average. The most common case of HSH includes two variables only; but larger subsets can also occur. HSH mostly means that the systematic variation of data cannot be captured sufficiently well by one factor only. The consequence is bad model fit.
In the framework of confirmatory factor analysis a bifactor model is described that includes a HSH factor besides a general factor. Such a model enables the capturing of systematic variation due to HSH so that bad model fit is avoided. Furthermore, the relationship between the two factors can be investigated. This is demonstrated in real data and by the results of a simulation study.

Oral presentations session title:

Structural Equation Modeling (SEM)

Title

A new approach for specifying composites in structural equation models: The Henseler– Ogasawara specification

Author(s)

Florian Schuberth

Univeristy of Twente, Enschede, Netherlands

Abstract

For a long time, it was not clear how to properly incorporate composites, i.e., (weighted) linear combinations of variables, in structural equation models. Although the literature on structural equation modeling provides various approaches such as the one-step approach and the two-step approach, they do not allow researchers to model composites in a flexible way as they are accustomed from modeling latent variables. To overcome this limitation, the recently introduced Henseler-Ogasawara (H–O) specification is presented. In the H–O specification, not only one composite, but as many composites as indicators are extracted per block. Additionally, the relationship between the composites and their indicators is expressed by loadings instead of weights. As a consequence, the H–O specification allows researchers to model composites in a flexible way. Particularly, composites can be included as dependent variables in a structural model and composites with fixed or free weights can be specified.

Oral presentations session title:

Structural Equation Modeling (SEM)

2.14 Parallel sessions 1h30–13h00 Lecture room 1.2

Title

Proposing an Item Discrimination Index for the Tests that Select Top Students

Author(s)

Serkan Arikan[1], Eren Can Aybek[2]

[1]Bogazici University, Istanbul, Turkey; [2]Pamukkale University, Denizli, Turkey

Abstract

Item discrimination indices, such as item-total correlation, item-rest correlation, and the IRT a parameter, provide information about an item's capability to discriminate all students. However, there are tests that are used to select a small number of students. In the current study, a special case of Brennan's index (B10-90) is proposed to determine items that are useful to select a limited number of high-achieving examinees. Inspired by Brennan's formula, B10-90 is calculated as the difference between the performance of the upper 10 and lower 90 percent of the group on the item. The research question is "Which item discrimination index provides accurate information for tests used to select a small number of high-achieving examinees?"
Overall, 18 datasets (2x3x3) were generated on the following conditions: test length, ability distributions and item difficulty. Data generation and related analysis were done via R.
The first part of the results showed that the IRT a' parameter (the slope of the ICC at a point where the upper 10th and lower 90th percentiles intersect) had the highest correlation with B10-90 (r=0.93) and phi10-90 (r=0.92). In the second part, the congruence of an item discrimination index with percentage correct response curves was evaluated following the procedure for the graphical approach of the ETS. B10-90 provided systematically accurate information. As a result of both correlational and graphical evaluations, B10-90 is recommended for tests when the purpose is to select a limited number of high-achieving examinees.

Oral presentations session title:

Item Response Theory (IRT) and DIF

Title

Using response times to study between-country comparability in large-scale educational assessment

Author(s)

Jesper Tijmstra[1], Maria Bolsinova[1], Leslie Rutkowski[2,3], David Rutkowski[2,3]

[1]Tilburg University, Tilburg, Netherlands; [2]Indiana University, Bloomington, USA; [3]Centre for Educational Measurement, Oslo, Norway

Abstract

In large-scale educational assessment, establishing the comparability of country-level outcomes is of crucial importance. This is often approached by considering between-country measurement invariance for the measurement model for response accuracy, and studying whether there may be differential item functioning at this level, with for example an item being relatively easy for one country compared to the international population. However, for outcomes to be comparable, we need to be confident that the response processes are comparable across countries, and hence we need to carefully study whether there are between-country differences in how respondents deal with each of the items on the test.

Process data provide a highly valuable source of information for studying possible between-country differences in how respondents take the test. We propose the use of response times for studying response process comparability, which allows for a much richer picture than is possible using standard differential item functioning analysis (which only considers response accuracy).

We will discuss three ways in which response times can be used to provide evidence of (in-)comparability of the response processes across countries: (1) a contrast of the item-level distribution of response time; (2) a contrast of the proportion and nature of disengaged responses to each item; (3) a test-level assessment of the relationship between speed and ability in each country, based on the hierarchical model for response time and accuracy. Our application to PISA 2015 Science data shows relevant country differences for all three aspects, also on items not originally flagged as showing DIF.

Oral presentations session title:

Item Response Theory (IRT) and DIF

Title

Generalizing beyond the test: Permutation-based profile analysis for explaining DIF using item features

Author(s)

Maria Bolsinova[1], Jesper Tijmstra[1], Leslie Rutkowski[2], David Rutkowski[2]

[1]Tilburg University, Tilburg, Netherlands; [2]Indiana University, Bloomington, USA

Abstract

Profile analysis (Verhelst, 2012) is one of the main tools for studying whether differential item functioning (DIF) of items on a test can be related to specific features of the items on the test (e.g., their response format, content domain, or the depth of knowledge required to answer the item). While relevant, profile analysis in its current form has two restrictions that limit its applicability in practice: It assumes that all test items have equal discriminations, and it does not test whether conclusions about the item feature effects generalize outside of the test (i.e., whether they hold in general rather than just for the specific set of items on the test). This presentation addresses both of these limitations, by generalizing profile analysis to work under the two-parameter logistic model and by proposing a permutation test that allows for generalizable conclusions about item-feature effects. This latter step is especially important in practice, where test developers will be interested in determining whether item-feature effects can be expected to hold for similar future tests as well, or whether observed patterns in DIF should not be expected to generalize beyond the particular testing setting. The developed methods are illustrated using PISA 2015 Science data, which show important differences between the results obtained using 'standard' profile analysis versus those obtained using its generalizable extension. The methods are also evaluated using simulation studies that assess their Type I error rate and power.

Oral presentations session title:

Item Response Theory (IRT) and DIF

Title

Pervasive Differential Item Functioning

Author(s)

Paul De Boeck[1], William Goette[2]

[1]Ohio State University, Columbus OH, USA; [2]University of Texas Southwestern Medical Center, Dallas TX, USA

Abstract

The stream of methodological studies of differential item functioning (DIF) is unstoppable. One reason is that DIF detection is very important. Another reason is that DIF cannot be identified without implicit or explicit assumptions. We believe that most approaches minimize DIF detection (and detection of measurement invariance violations), including the regularization methods and factor model alignment methods. In the presence of DIF that pervades a large part of the test, minimizing DIF detection can have far reaching consequences for group mean differences. A possible cause of pervasive DIF is that one or more item covariates (item properties) affect the item difficulties (intercepts) in a way that depends on group membership. This can be the cause of group mean differences. The principle is explained by De Boeck and Cho (2021). We will present: 1. A simulation study showing that pervasive DIF can create group mean differences while no substantial DIF is detected. Traditional DIF methods and measurement invariance methods fail to detect pervasive DIF that causes group mean differences. 2. An outline for the detection of pervasive DIF. 3. Results from a pervasive DIF analysis using data from the Boston Naming Test (BNT; Kaplan, et al., 1983). The BNT is the most commonly used language test among neuropsychologists and a critical diagnostic step in detecting and characterizing neurocognitive disorders. A common finding is that there are group mean differences between Caucasians and African Americans. The research question is whether not pervasive DIF can explain these group mean differences.

Oral presentations session title:

Item Response Theory (IRT) and DIF

2.15 Parallel sessions 1h30–13h00 Lecture room 1.3

Title

A new perspective on Cramer's phi

Author(s)

Szymon Czarnik

Jagiellonian University, Krakow, Poland

Abstract

Cramer's phi coefficient is likely the most often used chi-square based measure of association between nominal variables. In various sources, it is labeled as an asymmetric measure with desired property of being normalized to the minimum of 0 and the maximum of 1 yet being hard to interpret in the case of intermediate values. We propose that it would be more apt to construe Cramer's phi as an asymmetric measure and provide a straightforward probabilistic interpretation of its values.

Oral presentations session title:

Statistics

Title

The practice of power analysis and Its implications: A meta-science review

Author(s)

Kathryn Hoisington-Shaw[1], Octavia Wong[2], Zhaojun Li[1], Mark Matthews[1], Duane Wegener[1], Jolynn Pek[1]

[1]The Ohio State University, Columbus, OH, USA; [2]York University, Toronto, Canada

Abstract

Power analysis is considered the gold standard to justifying sample size for a research study, but the limits of its application continue to be misunderstood in practice. We report on survey results indicating that psychologists (incorrectly) consider power relevant to interpreting a statistically significant result, and this belief covaries with where they learn about statistical power. The link between limited access to learning about power and a common misconception about power analysis provides a context for our meta-science review on the practice of power analysis in Psychological Science papers published in 2017 and 2021. Our review indicated that researchers are increasingly justifying sample size, though not necessarily through use of power analysis. However, they are also increasingly misapplying power analysis in the post-study phase. By specifying power, sample size, and Type I error, researchers have increased in calculating a minimum detectable effect size after data collection. Though not generally explicitly stated, the idea seems to be that a sample is sufficient if the obtained effect is larger than the minimum detectable effect. However, depending on the population effect size, simulation data show that effects smaller than the minimum detectable can often be significant. In the context of our meta-science review, we highlight challenges, misconceptions, and misapplications concerning power analysis and discuss other practices that can improve research.

Oral presentations session title:

Statistics

Title

A user-friendly tool to code planned comparisons for statistical analyses.

Author(s)

Umberto Granziol[1], Maximilian Rabe[2], Andrea Spoto[1], Giulio Vidotto[1]

[1]Department of General Psychology, University of Padova., Padova, Italy; [2]Department of Psychology, University of Potsdam, Potsdam, Germany

Abstract

Prior to statistical analyses, scientists and researchers often have specific hypotheses about the difference between mean groups. Their hypotheses are often tested by post-hoc comparisons with the main statistically significant and/or interaction effects. With the exception of exploratory studies, post-hoc comparisons can increase Type I error rates and reduce statistical power. A well-known solution is to use planned comparisons. Nevertheless, it is difficult to understand and implement them, especially the customized comparisons for interaction effects. The purpose of this presentation is to introduce appRiori, a Shiny application coded in R. AppRiori lets users upload research variables and guide them to the best comparison set that fits the hypotheses, both for the main effects and for the interaction. Using graphic explanations and reproducible data empirical examples, we proved that it was possible to understand the logic of planned comparisons and their interpretation when tested in a model. AppRiori can program in R all default planned comparison. In addition, via plug-and-play logic, users can also customize such comparisons even on interaction effects. Finally, it is possible to obtain R codes related to the planned comparison. The R code can be copied and applied directly before running statistical models such as variance analysis, linear (mixed) and nonparametric models. The impact and use of appRiori is discussed.

Oral presentations session title:

Statistics

Title

The effects of multicollinearity when testing congruence hypotheses in response surface analysis

Author(s)

Florian Scharf[1], Dominik Strutz[2], Johanna Paping[3], Simon Grund[3], Sarah Humberg[4]

[1]Kassel University, Kassel, Germany; [2]Chemnitz University, Chemnitz, Germany; [3]Hamburg University, Hamburg, Germany; [4]WWU Münster, Münster, Germany

Abstract

Response Surface Analysis (RSA) is often used to test congruence hypotheses, that is, if agreement between two sources of information (e.g., self-view vs. objective psychometric tests) is beneficial for an outcome variable. High multicollinearity is a common concern in RSA models and can arise due to the polynomial regression model underlying RSA and due to substantially correlated predictor variables. Unlike classic multiple regression, RSA relies on a set of auxiliary parameters which are computed from the regression weights and may behave considerably different from the regression weights under high predictor correlations. The aims of the present article are two-fold: (1) We investigated the behavior of the auxiliary RSA parameters and the test of the congruence hypothesis analytically and in a simulation study. (2) We investigated whether the negative effects of multicollinearity can be mitigated by regularization techniques such as ridge regression. In contrast to findings for traditional multiple regression, we found that the effects of multicollinearity on RSA depend on the sign of the predictor correlation. Classic regularization techniques such as ridge cannot be recommended for RSA models.

Oral presentations session title:

Statistics

2.16 Parallel sessions 16h00–17h30 Auditorium 1

Symposium Overview

Programs evaluation: Methodological quality and effect size estimation

Author(s)

Salvador Chacón-Moscoso[1,2], Susana Sanduvete-Chaves[1]

[1]Universidad de Sevilla, Seville, Spain; [2]Universidad Autónoma de Chile, Santiago, Chile

Abstract

It is usual to find methodological weakness in the evaluation of intervention programs in the different fields of application (sports, organizational, health or clinical, among others). The fundamental aspects of how the interventions are carried out are not always in a high level of quality. Additionally, the problems are not always solved in the most effective ways. This hinders the integrated accumulation of knowledge and, therefore, the growth of science and its applicability to technology. In this symposium, we present the advances of our research group 'Methodological innovations in programs evaluation' (HUM-649, Junta de Andalucía) in these issues, together with other research groups from Europe and America. Four oral presentations are included. The first and the second ones show validity evidence of two scales to measure the methodological quality, in experimental and quasi-experimental studies, and in observational studies (with low level of intervention), respectively. The third presentation is an invariance study of a scale to measure work climate in emergency departments. Finally, the last presentation, based on previous methodological quality scales, is a meta-analysis about the effectiveness of psychological therapies for patients with chronic pain. With these presentations, we provide a possible approximation to support evidence of decision-making process in programs evaluation implementation. Specifically, we offer two tools to measure methodological quality for studies with any kind of design, another tool to measure work climate in emergency departments for different kind of hospitals and different countries, and conclusions about the effectiveness of psychological interventions and its relationship with methodological quality.

Title

Methodological Quality Scale: Convergent – discriminant validity evidence

Author(s)

José Mena-Raposo[1], Salvador Chacón-Moscoso[1,2], Susana Sanduvete-Chaves[1], Daniel López-Arenas[1]

[1]Universidad de Sevilla, Seville, Spain; [2]Universidad Autónoma de Chile, Santiago, Chile

Abstract

The 10-item Methodological Quality Scale (MQS) presents clear inclusion criteria for item selection. It is easy to be used; it requires only 2-4 minutes to be completed. It is highly flexible, as it is applicable to both randomized and non-randomized controlled trials. Additionally, it presents adequate levels of reliability and validity evidence. This study examined its convergent and discriminant validity. It was contrasted with Black and Down's tool, PeDRO and RoB-2, which are representative of methodological quality constructs in terms of both internal and external validity. Fifty one studies on training programs for workers in organizations were randomly selected and coded with the four different scales by two experts previously trained. The general analysis involved inspecting the scale results and determining their metric properties. We studied, intercoder reliability; descriptive analysis of items (mean, median, standard deviation, kurtosis, skewness); dimensionality of each scale, and metric characteristics of each factor (mean, variability, McDonald's Omega and average discrimination); and correlation analysis to assess the relationship between dimensions. Finally, multitrait-multimethod matrix was applied to obtained convergent and discriminant validity evidence. The results showed that our scale has good convergent validity, demonstrating a strong correlation with other established quality scales. These findings support the use of the proposed MQS as a reliable and valid tool for assessing the quality of experimental and quasi-experimental research studies, particularly those aimed at interventions in health sciences, behavioral sciences and, in general, social sciences, where the use of Randomized Control Trials (RCT) is not always possible.

Symposium title

Programs evaluation: Methodological quality and effect size estimation

Title

Methodological Quality Scale for studies based on Observational Methodology (MQSOM)

Author(s)

Daniel López-Arenas[1], Susana Sanduvete-Chaves[1], Salvador Chacón-Moscoso[1,2], M.Teresa Anguera[3]

[1]Universidad de Sevilla, Seville, Spain; [2]Universidad Autónoma de Chile, Santiago, Chile; [3]University of Barcelona, Barcelona, Spain

Abstract

No existing instrument addresses the methodological quality in studies based on observational methodology. Consequently, there is a lack of framework to guide both professionals along their practice in this type of study and journal reviewers when assessing their acceptance or rejection. This work studies the psychometric properties of the items from a Methodological Quality Checklist for studies based on Observational Methodology (MQCOM; Chacón-Moscoso et al., 2019) to validate a Methodological Quality Scale providing evidence about its internal structure in terms of validity and reliability, and about the discriminant power of each item. Six hundred and fifty journal articles, which applied observational methodology, were obtained through the databases PsycINFO, SportDiscus, Psicodoc, SCOPUS and Web of Science and after asking experts in observational methodology about primary studies, published or not. Two independent researchers codified these articles. After obtaining the intraclass correlation index both for inter and intra-coder reliability, an Exploratory Factor Analysis was carried out with half of the sample to explore the factor structure of the scale and, afterwards, a Confirmatory Factor Analysis was carried out with the other half of the sample to confirm the structure previously obtained. The number of dimensions and the concrete items that formed each dimension with adequate psychometric properties were obtained. This study developed an useful scale to enhance the methodological quality of studies based on observational methodology. This original contribution could become a milestone in the development of a methodological culture on systematic observation.
Funding: PID2020-115486GB-I00, 2021 SGR 00718, EXP_74847

Symposium title

Programs evaluation: Methodological quality and effect size estimation

Title

Work climate scale in emergency departments: A measurement invariance study

Author(s)

José A. Lozano-Lozano[1], Susana Sanduvete-Chaves[2], Salvador Chacón-Moscoso[2,1], F. Pablo Holgado-Tello[3]

[1]Universidad Autónoma de Chile, Santiago, Chile; [2]Universidad de Sevilla, Seville, Spain; [3]Universidad Nacional de Educación a Distancia, Madrid, Spain

Abstract

The abridged version of the work climate scale (Lozano et al., 2021) is a 24-item instrument that has good metric properties. Confirmatory factor analysis yielded appropriate global fit indices with a Chilean sample, X2(248) = 367.84; p < 0.01, RMSEA = 0.06, CI 90% [0.05, 0.07], SRMR = 0.08, GFI = 0.9, AGFI = 0.96, CFI = 0.98, NFI = 0.95 and NNFI = 0.98; along with test criteria validity, XY = 0.68, p < 0.001; and excellent reliability, α = 0.94 and = 0.94. This work presents an empirical study of measurement invariance in the area of work environment in emergency departments across countries (Chile and Spain) and gender (men and women). Health professionals from the emergency department fulfilled the scale. In Chile, they were 258 participants, 61.2% women, aged 20 to 64 years (M = 31.95, SD = 7.30). In Spain, they were 307 participants, 64.5% women, aged between 23 and 65 years (M = 40.70, SD = 8.71). Criterion-related validity evidence was examined through the working group survey (Perry et al., 2005). Of the models tested, an Exploratory Model of Structural Equations (ESEM) fits the data best. Measurement invariance was tested with Multi-Group Confirmatory Factor Analysis (MG CFA). In general, MG CFA supported the measurement invariance of the scale in all groups. Some correlational differences arisen between countries are discussed.
Funding: This research was supported by the grant PID2020-115486GB-I00, Ministry of Science and Innovation –Ministerio de Ciencia e Innovación–, MCIN/AEI/ 10.13039/501100011033, Government of Spain, Spain.

Symposium title

Programs evaluation: Methodological quality and effect size estimation

Title

Effectiveness of psychological interventions to decrease cognitive fusion in patients with chronic pain: a systematic review and meta-analysis

Author(s)

Susana Sanduvete-Chaves[1], Salvador Chacón-Moscoso[1,2], Francisco J. Cano-García[1]

[1]Universidad de Sevilla, Seville, Spain; [2]Universidad Autónoma de Chile, Santiago, Chile

Abstract

The specific role of cognitive fusion in chronic pain, one of the six key components of the psychological flexibility model, has not yet been well established. The aim was to gauge the effectiveness of psychological interventions to decrease cognitive fusion (PROSPERO-CRD42021255028). The Web of Science, SCOPUS, Medline, and PsycInfo databases were searched for relevant primary studies done up to January 2022. The inclusion criteria for the studies were patients with a chronic pain diagnosis, psychological interventions for patients, and measurements of cognitive fusion. The risk of bias of the selected studies was measured using a methodological quality scale and the average effect sizes (Hedges' g) were calculated. This review included 12 papers. According to the findings, cognitive fusion decreased significantly after the intervention. The effect sizes were small/medium in the post-test, g = -0.333, $p < .001$, 95% CI [-0.478, -0.187]; and medium in the long-term follow-up, g = -0.534, $p < .001$, 95% CI [−0.759, -0.309]. A similar tendency was found for studies with RCTs in post-test, g = -0.622, p = .042, 95% CI [−1.221, -0.023] and short-term follow-up, g = -0.793, $p < .001$, 95% CI [−1.176, -0.401]. Moderator variables such as unemployment, gender, pain intensity, level of depression before the intervention, and duration and modality of the intervention were identified.
Funding: This research was supported by the grant PID2020-115486GB-I00, Ministry of Science and Innovation, MCIN/AEI/ 10.13039/501100011033, Government of Spain, Spain; and the Andalusian Plan for Research, Development and Innovation (PAIDI 2020), Government of Andalusia, Spain [ref. PY20-01122].

Symposium title

Programs evaluation: Methodological quality and effect size estimation

2.17 Parallel sessions 16h00–17h30 Auditorium 2

Symposium Overview

Interdisciplinary research methodology - moving forwards

Author(s)

Hilde Tobi

Wageningen University & Research, Wageningen, Netherlands

Abstract

Generally speaking, policy, academia and society call for an increase in interdisciplinarity to study complex issues such as the sustainability of care systems and consequences of the climate crisis. The title of this fourth EAM symposium on interdisciplinary research methodology reflects development in mixed methods and qualitative interdisciplinary research methodology. In this symposium we'll present an array of recent developments in interdisciplinary research methodology that gives food for thought on research integrity, ethics, and contributes to the further growth of the interdisciplinary research methodology knowledge base.

Title

Applied phenomenological analysis according to Giorgi. An interdisciplinary analysis of the distance caregiving triad's demands on two levels within the study 'ROAD – CaRegiving frOm A Distance'

Author(s)

Farina Buenning[1], Vivian Lou[2], Huanran Liu[2], Andrea Budnick[1]

[1]Institute of Medical Sociology and Rehabilitation Science, Charité – Universitätsmedizin Berlin, Berlin, Germany; [2]Department of Social Work & Social Administration, Sau Po Centre on Ageing, The University of Hong Kong, Hong Kong, Hong Kong

Abstract

Introduction: Social changes, such as professional and personal mobility, lower the potential of local home support for care receivers and might increase distance caregiving arrangements. To generate recommendations for reliable care arrangements, we need to know the interpersonal and structural demands of the distance caregiving triad (care receiver, distance caregiver, informal/professional local network).
Methodology: Our interdisciplinary team applies a qualitative multi-method design in 20 triads, including sociological and gerontological approaches, as well as perspectives of nursing science and social work. We conduct guided interviews with care receivers (n=20), distance caregivers (n=20), and the local network (n=20), and participatory observations in care receivers' homes (n=20). We apply phenomenological analysis according to Giorgi on two levels: (1) in analyzing the actors' demands within every triad and (2) in comparing the triads regarding to sociodemographic characteristics. Moreover, we triangulate the results of the guided interviews and the participatory observations.
Results: We find differences in the actors' perceived autonomy and in their interpersonal expectations. We are currently analyzing congruencies and discrepancies intergenerationally and actor-specifically, as well as differences between triads with different sociodemographic backgrounds.
Discussion: The cooperation in the international, interdisciplinary team is both, helpful in composing comprehensiverecommendations, and challenging in combining all approaches. Study results contribute to needs-based support services for higher reliability in distance caregiving arrangements. In doing so, the interpersonal approach includes perspectives of allactors: while care receivers can stay in their homes as long as possible, distance caregivers and the local network get more safety.

Symposium title

Interdisciplinary research methodology - moving forwards

Title

The validity of concept mapping in interdisciplinary research

Author(s)

Jarl Kampen[1], Jos Hageman[1], Marian Breuer[2], Hilde Tobi[1]

[1]Wageningen University & Research, Wageningen, Netherlands; [2]Radboud University Medical Center, Nijmegen, Netherlands

Abstract

Concept mapping is a technique used to visually organize and represent knowledge or information. It starts with the elicitation of statements and ends in a diagram or map that shows the relationships between concepts or ideas. Concept Mapping is a mixed-methods technique that seemed promising for interdisciplinary and intercultural research where concepts are owned by people with very different backgrounds. Therefore, we tried to dig deep into the procedure and detected a number of issues related to transparency, theory and multi-variate statistics. In this contribution, we'll present the findings of our dig and the results from a simulation study.

Symposium title

Interdisciplinary research methodology - moving forwards

Title

Inclusive methodologies: interdisciplinary challenges and actionable methods to improve inclusive research

Author(s)

Leen Sterckx[1], Maria Luce Sijpenhof[2]

[1]The Netherlands Institute For Social Research, The Hague, Netherlands; [2]The Netherlands Institute for Social Research, The Hague, Netherlands

Abstract

The Netherlands Institute for Social Research is an independent institute that aims to better policy through interdisciplinary scientific knowledge. To do so, the institute has a responsibility to conduct inclusive research and to provide accurate and complete information on citizens' perspectives. In the process, importance is attached to researchers' objectivity. However, this seems unattainable, as critical paradigms show that dominant assumptions of objectivity should be challenged. The institute's use of seemingly neutral description and categorizations are insufficient for truly inclusive research. Improvement requires both researchers' reflexivity on their attitudes, assumptions, biases, (normative) concepts and actionable methods to increase participation in research. We designed a multi phased project to study the ways in which methodology may be improved to include a variety of underrepresented groups in the Institutes research. The first subproject will focus on ethnic, religious and racial minorities. Using a critical approach, this subproject includes literature review, a discursive analysis of internal documents and procedures, expert interviews and focus groups with the relevant minority groups. Our goal is to develop specific inclusive instruments, methods and protocols tailored to the needs of the Institute. In the process we notice that the various epistemological and theoretical disciplines are not easily reconciled. In our presentation we provide an account of the discussions we encounter and how we attempt to overcome challenges.

Symposium title

Interdisciplinary research methodology - moving forwards

Title

Complex research designs: methodological and ethical considerations in interdisciplinary, mixed methods research

Author(s)

Suzanne Roggeveen, Claudia Hartman, Joep Schaper

The Netherlands Institute for Social Research, The Hague, Netherlands

Abstract

Interdisciplinary and mixed methods research are often used to solve complex research puzzles. Looking through various theoretical lenses or using multiple methods can, at least in theory, create new kinds of empirical insights. However, these kinds of research are not always easy, especially when combined. Team members might work from dissimilar research paradigms, might have different power positions within the team and might not share the same norms and values. These challenges also bring methodological and ethical considerations that are particularly apparent within interdisciplinary mixed methods research. In this paper we describe some of the intersections between interdisciplinarity and mixed methods research in the social sciences and their methodological and ethical consequences. Without trying to be complete we will highlight epistemological and ontological challenges in different disciplines and research methods traditions that are usually discussed at the start of the research, integration of analysis in the middle of the project and synthesis in the writing up stage. In doing so we describe best practices as described in methodological literature and use examples from an interdisciplinary, mixed methods research project at the Netherlands Institute for Social Research (SCP). This project looks at COVID skepticism in The Netherlands through the lens of a survey, a digital ethnography, and a quantitative content analysis of social media data and newspaper articles. Reflections on this project showcase epistemological and ontological challenges in everyday research practice. Finally, we will address hopeful solutions for future research.

Symposium title

Interdisciplinary research methodology - moving forwards

2.18 Parallel sessions 16h00–17h30 Auditorium 3

Symposium Overview

New developments in modeling response times in psychological assessments

Author(s)

Augustin Mutak[1], Sören Much[2]

[1]Freie Universität Berlin, Berlin, Germany; [2]Martin-Luther-Universität Halle-Wittenberg, Halle, Germany

Abstract

With the rise of computerized testing, there has been a trend of increased availability of response time data in psychological assessments. This data has been used to improve our understanding of response processes and test-taking behavior, in particular by examining examinee's speed, non-solution behavior like guessing and omissions and time components for stimulus processing and responding. However, the interpretation of response time and omission data is not yet fully clear. The same observed events (such as a specific response time or an omitted item) can be caused by different mechanisms, which in turn imply different psychological interpretations of examinee's behavior. This symposium presents several models which provide tools for exploring potential causes of test-taking phenomena. The models which we present (a) examine examinee's persistence during a task solution and how their progress can account for partial guessing and omissions, (b) allow researchers to look into how examinees' behavior on previous items might relate to their behavior on the following items and (c) show how, with the help of eye-tracking data, an examinee's speed can be split into two underlying components that are more informative than the general speed. With these novel psychometric models, the estimation of item parameters can sometimes become very challenging. Thus, the symposium also presents (d) a new estimation algorithm based on deep learning which enhances parameter estimation in models with responses and response times.

Title

Deep Learning Approaches for Factor Analysis of Responses and Response Times

Author(s)

Rudolf Debelak[1], Christopher John Urban[2]

[1]Zürich, Zürich, Switzerland; [2]University of North Carolina at Chapel Hill, Chapel Hill, USA

Abstract

An important problem in the application of psychometric models is the selection of suitable algorithms for parameter estimation. In a recent publication, Urban and Bauer (Psychometrika 86:1-29, 2021) proposed an estimation algorithm based on deep learning for item parameter estimation for large sample sizes. We first give an overview on the principal ideas of this approach, and how it related to classical estimation methods for models of item response theory. Second, we will evaluate the accuracy of this approach for two types of factor models: a) a log-normal factor model for response times, b) a hybrid factor model for responses and response times, which is related to previously proposed methods for responses and response times. The results of the simulation studies indicate that the deep learning algorithm can be used for an accurate item parameter estimation for these models even in relatively small datasets and is computationally fast in large datasets. The evaluated algorithm is freely available in a Python package.

Symposium title

New developments in modeling response times in psychological assessments

Title

Looking into Time-on-Task: A Hierarchical Model with Multiple Time Components Applied to Eye-Movement Data

Author(s)

Tobias Deribo[1], Daniel Bengs[1], Frank Goldhammer[1,2], Ulf Kroehne[1]

[1]DIPF | Leibniz Institute for Research and Information in Education, Frankfurt a.M., Germany; [2]Centre for International Student Assessment (ZIB), Frankfurt a.M., Germany

Abstract

Jointly modeling latent speed and ability has been proven helpful in multiple applications (De Boeck & Jeon, 2019). Here, log-data-based time-on-task is commonly used to measure latent speed, possibly aggregating multiple parts of the underlying response process (e.g., Johnson-Laird, 1994). Process data derived from eye-tracking studies provides opportunities to identify and separate time components related to different aspects of task processing. Assuming that durations of fixation of different areas of interest indicate test takers‘ active engagement with them (Just & Carpenter 1980) and relate to distinct meaningful aspects of the response process, we extract time components related to fixating item stimulus and response options from eye-movement data. We propose an extension of the hierarchical model (van der Linden, 2007), that provides a joint model for the time component measures and response accuracy. The proposed model for latent ability, stimulus speed, and response speed is then compared to a unidimensional ability model and a model with only a general speed parameter. All analyses are based on eye-movement data from a non-verbal intelligence test (Weiß et al., 2006) taken by N = 186 university students (Kasneci et al., 2021). A Bayesian approach using Stan (Stan Development Team, 2020) was applied for estimation. The proposed models using two separate speed parameters showed higher predictive accuracy (Vehtari et al., 2016) and higher approximate relative efficiency (de la Torre & Patz, 2005) when compared to models with a general speed parameter. The results highlight how parallel information, like eye-movement data, can further inform existing psychometric models.

Symposium title

New developments in modeling response times in psychological assessments

Title

Modeling omissions in tests as dependent on previous test behavior

Author(s)

Augustin Mutak[1], Esther Ulitzsch[2], Sören Much[3], Jochen Ranger[3], Steffi Pohl[1]

[1]Freie Universität Berlin, Berlin, Germany; [2]IPN Kiel - Leibniz Institute for Science and Mathematics Education, Kiel, Germany; [3]Martin-Luther-Universität Halle-Wittenberg, Halle, Germany

Abstract

In order to adequately account for missing values in tests, it is essential to have a good understanding of how they emerge. Most of the current approaches place missing responses into the context of low ability, disengagement, or general test-wiseness. However, the mechanism with which missing data in psychological tests is produced is still not fully known, since it is likely that they are caused by multiple factors. There are little findings in the literature which reveal how behavior on previous items can be connected with omissions in subsequent items. However, previous behavior in the test, such as taking too much time or not performing well may impact test-takers strategy. To explore this, we develop a new model, which includes responses, response times and omissions on the manifest level and their respective latent constructs. In the model, we relax the assumption of conditional independence between responses or response times on an item and omissions in the subsequent item. By allowing for these residual correlations to be estimated, we explore if investing relatively too much time on one item can lead examinees to become more hesitant into investing time to solve the next item. We also investigate whether comparably worse performance on a previous item impacts the occurrence of missing values on further items. We conduct a simulation study to test our model performance. To illustrate its use, we apply it to an empirical dataset exploring whether behavior in responding to previous items may explain omissions in later items.

Symposium title

New developments in modeling response times in psychological assessments

Title

A ballistic accumulator model to account for examinee's persistence and partial knowledge guessing

Author(s)

Sören Much[1], Jochen Ranger[1], Augustin Mutak[2], Steffi Pohl[2]

[1]Martin-Luther-Universität, Halle, Germany; [2]Freie Universität, Berlin, Germany

Abstract

When test-takers are under time pressure or if a test is of lower importance for them, it can be reasonable to put less effort in the solution of a test item, take an educated guess and move on to the next question. This phenomenon has often been modeled as a dichotomy of rapid guessing versus non-rapid solution behavior. We propose a model that allows for a more nuanced view of test-taking behavior, taking into account that test-takers can solve parts of an item and then reach a response decision based on the current status of their solution. The model is a development of the Linear Ballistic Accumulator model (LBA, Brown & Heathcote, 2008). The LBA describes a race of information accumulators that represent the response options towards a common response threshold. This model can be extended with a time-based accumulator that stops the information accumulators once it reaches the threshold and reflects a premature end of solution efforts. The values of the information accumulators at the time of stopping then yield the probabilities of a correct or wrong response or even omissions. As it has been done for other sequential sampling models (e.g., van der Maas et al., 2011), we introduce person and item parameters into the model, present ways to estimate them and apply the model to empirical data.

Symposium title

New developments in modeling response times in psychological assessments

2.19 Parallel sessions 16h00–17h30 Auditorium 4

Title

How Do We Know that a Bifactor Model is Optimal? An Example of Model Validation Using the BESS TRS-P Norm Dataset

Author(s)

Christine DiStefano, Jungsun Go

University of South Carolina, Columbia, USA

Abstract

When examining if data fit a hypothesized measurement model, it is common for researchers to evaluate competing models, where models are constructed based on alternative theoretical perspectives. Evaluating multiple models supports a strong program of validity evidence. While models may be compared on multiple aspects (e.g., theoretical, statistical, and interpretive criteria), it may be argued that fit information, especially global fit indices, are heavily considered when selecting among competing models. When examining fit indices, characteristics of the tested model may impact the fit indices provided. For example, the bifactor model has received a lot of attention as a modeling strategy that may be preferred on the basis of fit index information as this model often illustrates favorable fit indices due to a high number of paths being estimated.

In recent years, use of the bifactor model as an alternative has increased dramatically. Bifactor models may be preferred over other models in situations where a general factor (relating to all items) is present and one or more group factors (relating to subset of items) which are separate from the general domain However, some researchers have cautioned against use of the bifactor model due to difficulties with interpretation and model "overfitting". To address such concerns, additional indices have been proposed to aid in selection when a bifactor model is present. However, these indices may not be frequently included in studies evaluating a test's structure. This study will examine these indices and illustrate their use with an empirical example.

Oral presentations session title:

Bifactor and MTMM models

Title

On the Performance of Different Regularization Methods in Bifactor-(S-1) Models with Explanatory Variables—Caveats, Recommendations, and Future Directions

Author(s)

Benedikt Friemelt[1], Christian Bloszies[1], Maximilian S. Ernst[2], Aaron Peikert[2], Andreas M. Brandmaier[2], Tobias Koch[1]

[1]Friedrich Schiller University Jena, Jena, Germany; [2]Max Planck Institute for Human Development, Berlin, Germany

Abstract

Regularization methods in linear regression models with manifest variables have been shown to be effective in selecting key predictors from a set of many variables, while improving predictions for novel observations. Regularization methods are particularly attractive for the analysis of complex multidimensional data when theory development is the primary goal; for example when researchers attempt to predict general or specific factors in bifactor models using many potentially relevant predictors. However, applications of regularization methods in such models are still scarce. In a simulation study, we examined the performance of different regularization methods in bifactor-(S-1) models, varying the number of predictors, the correlations with the outcome (effect size), the underlying structure of multicollinearity as well as the sample size, the type of penalty, and a single-step versus a two-step approach. We explore potential caveats in the use of regularization methods in bifactor-(S-1) models, provide practical recommendations, and discuss future directions.
The presentation is based on a paper with the same name that has been published last December in Structural Equation Modeling: A Multidisciplinary Journal (https://doi.org/10.1080/10705511.2022

Oral presentations session title:

Bifactor and MTMM models

Title

Analysis of Assessment Dimensionality Using Multitrait-Multimethod Models in Rater Assessments

Author(s)

Denis Federiakin, Dominik Braunheim, Marie-Theres Nagel

Johannes Gutenberg University, Mainz, Germany

Abstract

The issue of assessment dimensionality is one of the most fundamental issues in psychometric research. However, the research on the analysis of dimensionality in assessments with raters is poor yet. The dominating paradigm for the analysis of rater assessments is Multifaceted Rasch Modeling from Item Response Theory (Eckes, 2009), where unidimensional models overwhelmingly dominate. In this presentation, we investigate Multi-Trait Multi-Method Factor Analysis models for rater assessments (Nussbeck et al., 2009), focusing on the analysis of assessment dimensionality. Based on analysis of real multidimensional data, we show that due to the constraints on factor loadings dedicated to extracting criteria-relevant variance (not virtual-item-relevant variance; Robitzsch & Steinfeld, 2018) in the construct-relevant space, many potentially distinct models become equivalent Schmid-Leiman (1957) transformations of each other (Gignac, 2016). Particularly, when conceptualizing criteria, dimensions, and the general factor as higher-order or bifactor structures, Schmid-Leiman transformations become possible as long as all items from the same parenting criteria have the same value of factor loadings in the construct-relevant space of person parameters. We show how a total of 15 potentially conceivable structures can be organized into 9 distinct models with differing model assumptions and model fit. We provide a real data example of Critical Online Reasoning (COR) assessment (Zlatkin-Troitschanskaia et al., 2021), where student performance in online information processing was scored by 3 raters on 6 different criteria measuring 3 dimensions of the COR construct. Based on this analysis, we discuss limitations and potential of Multi-Trait Multi-Method factor analysis models for rater assessments.

Oral presentations session title:

Bifactor and MTMM models

Title

The Generalized Total Entropy Fit Index for Bifactor Structures with correlated general factors

Author(s)

Hudson Golino[1], Marcos Jimenez[2], Alexander Christensen[3], Luis Garrido[4]

[1]University of Virginia, Charlottesville, USA; [2]Universidad Autonoma de Madrid, Madrid, Spain; [3]Vanderbild University, Nashville, USA; [4]Ponticia Universidad Catolica Madre y Maestra, Santo Domingo, Dominican Republic

Abstract

Exploratory Graph Analysis is a subfield of network psychometrics devoted to using network methods for dimensionality, item, and scale analysis. Recently, a hierarchical EGA method was proposed and showed higher accuracy than traditional methods in estimating the dimensionality structure of data generated using a bifactor model with correlated general factors. However, there is no fit metric currently available that can be used to investigate the fit of the dimensionality structure estimated via hierarchical EGA. The current presentation introduces the first fit measure for hierarchical EGA, and bifactor structures with correlated general factors termed the generalized total entropy fit index (GenTEFI). GenTEFI is a generalization of the total entropy fit index developed by Golino et al. (2021), that uses metrics from quantum information theory to investigate the fit of dimensionality structures to the data. In this presentation, the results of a Monte-Carlo simulation will be presented, investigating the accuracy of GenTEFI in detecting the correct dimensionality structure and comparing it to traditional SEM measures of fit (RMSEA, CFI, and SRMR). The result suggests that GenTEFI is as accurate as the traditional fit measures but with higher accuracy when the number of general factors is overfactored. Implications of the findings and new directions will be discussed.

Oral presentations session title:

Bifactor and MTMM models

2.20 Parallel sessions 16h00–17h30 Lecture room 1.2

Title

Exploring the validity evidence of a comprehensive assessment of first-year university students.

Author(s)

Milagrosa Sánchez-Martín[1], Juan F. Luesia[1], Isabel Benítez[2,3]

[1]Department of Psychology, Universidad Loyola Andalucía, Dos Hermanas, Spain; [2]Department of Methodology of Behavioural Science, University of Granada, Granada, Spain; [3]Mind, Brain and Behaviour Research Centre (CIMCYC), Granada, Spain

Abstract

Current educational trends highlight the relevance of both cross-curricular competencies and achievement as part of the academic competencies needed in university students. However, more evidence is needed to shed light on how these competencies impact academic performance levels. For this purpose, we analyze validity evidence of a comprehensive assessment of academic competencies (CompassIn), consisting of four cognitive and 21 non-cognitive competencies, through both receiver operating characteristic (ROC) curve analysis and network analysis. A total of 610 students participated in the study. In order to identify the academic performance levels, the students' performance after first-year (first-year GPA) was classified into low, medium, and high. The results of the ROC analysis showed that the cognitive competencies included in the CompassIn adequately classified students, regardless of the level of academic achievement (low, medium, or higher achievers) (AUC .70). In addition, the results of the network analysis showed that, in low achievers, the competence of conscientiousness was shown to be the most relevant in the model, followed by organizational skills. On the other hand, in high achievers, extraversion or communication skills were the most relevant competencies in the network. Two main conclusions can be highlighted from these results. First, CompassIn is useful for classifying students regardless of their level of academic achievement. Secondly, promoting non-cognitive competencies, such as conscientiousness or organizational skills, could be helpful for students with greater training needs.

Oral presentations session title:

Scale Development and Validation

Title

Methods of Measuring the Skills Mismatch in the Human Capital Study

Author(s)

Marcin Kocór, Szymon Czarnik

Jagiellonian University, Cracow, Poland

Abstract

The analysis of the labour market in terms of available skills, and especially the skills mismatch, has become increasingly important due to potential problems caused by existing asymmetries. The study of the asymmetry of the Polish labor market in terms of human capital took place relatively late, only in the 1990s, and on a limited scale. It was not until 2008 that the Human Capital Study was undertaken, in which, based on international experience, a unique approach to measuring the skills mismatch was proposed. It is based on declarative measurement of skills in surveys of various labor market participants. In our article we show what this method is based on, discuss its advantages, but also present its limitations. We then show how this approach has evolved in subsequent years and how it has been adapted to measure different categories of skills gaining recognition by various institutions in diagnosing mismatches in the labor market.

Oral presentations session title:

Scale Development and Validation

Title

Sensitivity to Punishment and Reward as Dispositional Traits in Face Perception: A Multilevel Analysis

Author(s)

Juan Carlos Oliver-Rodríguez

Universitat Jaume I, Castellón, Spain

Abstract

Sensitivity to punishment and reward have been shown to be instrumental constructs to account for behavioral dispositions in natural contexts. Questionnaire, biochemical and neuroimaging measures have been providing supporting evidence. Physiological correlates have been obtained with magnetic resonance imaging or electroencephalography in resting state and experimental situations such as attentional paradigms. It is however of interest to further explore their performance implications in different task conditions. In the present study, influence of these personality measures in a facial perception task is explored. Fifty five participants completed a Sensitivity to Punishment and Reward Questionnaire and participated in a facial perception task while their EEG activity was recorded. Stimuli were nine male and female faces which varied in physical attractiveness on the basis of normative data. Participants performed a face attractiveness evaluation during the task, and affective arousal and valence evaluations afterwards. Results from a multilevel analysis of the amplitude of the Late Positive Component (LPC) of the Event-Related Potential replicated previous findings of its association with affective arousal, which was here additionally modulated by sensitivity to punishment measures. An interaction among both personality measures and electrode site was also observed. Implications will be discussed.

Oral presentations session title:

Scale Development and Validation

2.21 Parallel sessions 16h00–17h30 Lecture room 1.3

Title

Explaining why individual religiosity measurement is noninvariant across countries with multilevel structural equation modelling

Author(s)

Alisa Remizova

University of Cologne, Cologne, Germany

Abstract

Individual religiosity measures have been widely used to compare individuals and societies. However, the cross-country comparability of measures has often been questioned. Comparability is a prerequisite for meaningful analyses of religiosity across countries and depends on measurement invariance. It indicates that the same construct is measured in the same way in all countries, so respondents understand the corresponding survey questions similarly. If the measurement is noninvariant, factors other than religiosity may systematically influence individual scores, and cross-country comparisons may be untrustworthy. While previous studies showed that religiosity measures lack invariance, it remains to be explained why they produce non-comparable data. The current research aims to systematically explain cross-country noninvariance of religiosity measurement. Specifically, I focus on the question that asks respondents about their belonging to a religious denomination. I use the joint dataset of the World Values Survey and European Values Study and employ the multilevel structural equation modelling method that allows accounting for noninvariance in a theoretically driven way. I explain the noninvariance by the cross-country differences in religious composition, regulation of religion, religious taxes, secularisation, history of communism, and cultural background. I conclude with directions for improving religiosity measurement and recommendations for practical researchers using the WVS/EVS data.

Oral presentations session title:

Measurement Invariance

Title

MixML-SEM: A parsimonious approach for finding clusters of groups with equivalent structural relations in presence of measurement non-invariance

Author(s)

Hongwei Zhao[1], Jeroen Vermunt[2], Kim De Roover[1]

[1]KU Leuven, Leuven, Belgium; [2]Tilburg University, Tilburg, Netherlands

Abstract

Structural equation modeling (SEM) is commonly used to explore relationships between latent variables, such as beliefs and attitudes. However, comparing structural relations across a large number of groups, such as countries, can be challenging. Existing SEM approaches may fall short, especially when measurement non-invariance is present. In this project, we propose Mixture Multilevel SEM (MixML-SEM), a novel approach to comparing relationships between latent variables across many groups that gathers groups with the same structural relations in a cluster, while accounting for measurement non-invariance in a parsimonious way. Specifically, MixML-SEM captures measurement non-invariance using multilevel CFA and then estimates the structural relations and mixture clustering of the groups by means of the structural-after-measurement (SAM) approach. In this way, MixML-SEM ensures that the clustering is focused on structural relations and unaffected by differences in measurement. MixML-SEM is particularly useful when sample sizes per group are too small to estimate partially group-specific measurement models (e.g., by multigroup CFA). In this case, accounting for measurement non-invariance with random parameters is more accurate and efficient. We demonstrate the effectiveness of MixML-SEM through simulations and a real data example, showing that it outperforms existing mixture SEM approaches.

Oral presentations session title:

Measurement Invariance

Title

Estimating structural paths in a multigroup Actor-Partner Interdependence Model (APIM) with small samples

Author(s)

Emma Somer, Carl Falk, Milica Miočević

McGill University, Montreal, Canada

Abstract

The Actor-Partner Interdependence Model (APIM) is a popular model used across the social sciences to model dyadic relationships. Structural equation modeling (SEM) can be used to investigate the relationships between latent variables in an APIM framework, thus allowing researchers to model individuals' influence on themselves and their partners. Recently, the structural-after-measurement (SAM; Rosseel & Loh, 2022) approach has been proposed as an alternative to SEM to alleviate some of the issues associated with model misspecification and small samples for maximum likelihood estimation. However, the performance of SAM and factor score regression (FSR), another two-step approach, for the APIM is unknown. We conduct simulation studies to evaluate the performance of SAM and the bias avoiding method for a multigroup APIM in small samples, and we compare the methods to SEM using maximum likelihood. In particular, we examine the influence of partial measurement invariance, reliability of the items, number of indicators, and measurement model misspecification on the bias, efficiency, Type I error, power, and coverage of the path coefficients for an APIM in the presence of correlated residuals. We implement a novel identification strategy for SAM in multigroup models, and we propose extracting standard errors and confidence intervals under the SAM framework using bootstrapping. Preliminary findings suggest SAM outperforms the bias avoiding method and SEM in terms of coverage and Type I error rates, particularly under low reliability conditions.

Oral presentations session title:

Measurement Invariance

Title

Scaling Metric Measurement Invariance Models

Author(s)

Eric Klopp, Stefan Klößner

Saarland University, Saarbrücken, Germany

Abstract

In the literature on measurement invariance, there are considerations about the choice of referent indicators, arguing that a wrong choice of the referent indicator may have ramifications for invariance tests. However, besides the scaling method that requires a referent indicator, two other scaling methods do not require a referent indicator. Thus, the general question is if either the choice of the scaling method in general or the choice of a specific referent indicator affects metric invariance tests. This paper clarifies these questions. We provide examples and a formal account proving that neither the choice of the scaling method in general nor the choice of a particular referent indicator affects the value of the discrepancy function and, therefore, the test statistic, too, for both the test of metric invariance models with either full or partial metric invariance. The results rely on an appropriate application of the scaling restrictions, which can be phrased as a simple rule: "Apply the scaling restriction in one group only, the invariance restrictions do the rest!". Additionally, we develop formulas to calculate the degrees of freedom for the 2-difference test, comparing a metric measurement invariance model to the corresponding configural model. This formula provides the interesting result that it is impossible to test the metric invariance of the estimated loading of exactly one indicator because metric measurement invariance models aimed at doing so are equivalent to the configural model.

Oral presentations session title:

Measurement Invariance

2.22 Poster session 2 14h00–15h00

Title

An empirical assessment of preregistration practices and prevalence in psychology meta-analyses

Author(s)

Alejandro Sandoval-Lentisco[1], Tom E. Hardwicke[2], Ruben Lopez-Nicolas[1], Miriam Tortajada[1], Jose Lopez-Lopez[1], Eric-Jan Wagenmakers[3], Julio Sanchez-Meca[1]

[1]University of Murcia, Murcia, Spain; [2]University of Melbourne, Melbourne, Australia; [3]University of Amsterdam, Amsterdam, Netherlands

Abstract

As it occurs with empirical studies, carrying out a meta-analysis requires making decisions about multiple aspects—e.g., data collection, data analysis, or reporting of the results. This flexibility—often referred to as 'researcher degrees of freedom'—entails a risk of bias, since researchers might make decisions with outcomes that are more in line with their personal preferences. Preregistering a study protocol has been proposed as a tool for detecting and reducing outcome-dependent biases arising from these 'researcher degrees of freedom'. Nonetheless, the degree to which a preregistration minimizes the risk of bias depends on how clear and comprehensive that preregistration is and whether deviations from protocol are disclosed and justified. In this study we had three aims. First, to measure the prevalence of preregistration in all meta-analyses published in psychology in 2021. Second, to assess the coverage of preregistrations by examining the extent to which a randomly selected subset of preregistered meta-analyses clearly prespecify key meta-analytic decisions that may incur a risk of bias. Lastly, to assess the deviations from preregistrations by comparing the decisions that were specified in the protocol to what was reported in the final article—for the same randomly selected subset of preregistered meta-analyses. Where deviations were encountered, we checked whether they were explicitly acknowledged and justified in the final article. The implications of the results shown here will be discussed. Funding: Regional Program for the Promotion of Scientific and Technical Research of Excellence (2022) - Seneca Foundation (grant no. 22064/PI/22). Grant PID2019-104080GB-I00 funded by MCIN/AEI/10.13039/501100011033.

Title

Properties of informative hypothesis tests and their integration into EffectLiteR

Author(s)

Caroline Keck[1], Axel Mayer[2], Yves Rosseel[1]

[1]Ghent University, Ghent, Belgium; [2]Bielefeld University, Bielefeld, Germany

Abstract

Within the framework of constrained statistical inference, we can test informative hypotheses. In these hypotheses, regression coefficients can be constrained to have a certain direction (for example $H_a : \beta_1 > 0, \beta_2 > 0, \beta_3 < 0$) or be in a specific order (for example $H_a : \beta_1 > \beta_2 > \beta_3$). Modified versions of the regular Wald, Score and likelihood-ratio test (LRT) as well as the distance statistic (D-statistic) can be used. These statistics are already well described in the literature, but practical information is largely missing. By means of simulation studies in the context of linear as well as generalized linear regression, we provide applied researchers with useful guidelines concerning type I and type II errors as well as regarding the choice of the informative test statistic.
Furthermore, research is often not only interested in inference concerning regression coefficients, but also regarding effects of interest. These effects may be average or conditional treatment effects, which are defined as a linear or non-linear combination of regression coefficients. The EffectLiteR approach provides a framework and R package for the estimation of average and conditional effects of a discrete treatment variable on a continuous outcome variable, conditioning on categorical and continuous covariates. We demonstrate how to integrate informative hypothesis testing into the EffectLiteR framework in the context of linear regression, while taking into account the stochastic nature of group sizes.

Title

Longitudinal measurement invariance assessment at the item level with Rasch family models: performances of the ROSALI algorithm for response shift detection

Author(s)

Myriam Blanchin, Yseulys Dubuy, Priscilla Brisson, Karima Hammas, Odile Stahl, Véronique Sébille

Nantes Université, Université de Tours, INSERM, MethodS in Patient-centered outcomes and HEalth Research, SPHERE, Nantes, France

Abstract

Background
In longitudinal Patient-Reported Outcome (PRO) studies, lack of longitudinal measurement invariance is a concern as it can bias the estimation of PRO change. This phenomenon theorized as response shift (RS) in PRO research is also viewed as a meaningful effect to quantify for a better understanding of patients' adaptation.
The "RespOnse Shift ALgorithm at Item level" (ROSALI) allows for item-level RS analyses using Rasch models assuming homogeneity of RS. However, RS effects may differ depending on clinical or psychological characteristics of patients. Hence, ROSALI was extended to take into account the effects of covariates on RS and PRO change. A simulation study was performed to evaluate its performances.
Methods
Different cases were evaluated according to: the i/ absence or presence of RS, ii/ type and size of RS, iii/ degree of RS heterogeneity. The performances of ROSALI were assessed regarding the rate of false RS detection (no simulated RS), the rate of correct RS detection (simulated RS), the accuracy of RS detection (affected groups and items) and PRO change bias.
Results
Rates of false detection of RS are low (between 2% and 7%). Rates of correct detection of RS are high (between 77% and 90%) when RS affects all response categories of an item similarly. The performances of ROSALI depend mainly on the sample size and the degree of RS heterogeneity.
Discussion
ROSALI satisfactorily prevents from mistakenly inferring RS. ROSALI is able to detect RS and identify the item and groups affected by RS especially as RS is homogeneous.

Title

Cognitive diagnostic computerized adaptive testing (CD-CAT) for classroom-level assessments

Author(s)

Pablo Nájera[1], Francisco J. Abad[1], Chia-Yi Chiu[2], Miguel A. Sorrel[1]

[1]Universidad Autónoma de Madrid, Madrid, Spain; [2]University of Minnesota Twin Cities, Minneapolis, USA

Abstract

Cognitive diagnosis modeling (CDM) has been posed as a promising statistical tool to evaluate the knowledge state of students, given its ability to identify the mastery or non-mastery of cognitive attributes. There are, however, two main obstacles that prevent CDM to be applied in real educational settings. First, parametric CDM usually require large sample sizes to provide accurate parameter estimates. Second, the formative nature of CDM assessments calls for a continuous and recurring set of evaluations throughout the course, which might require too much time from the classes. The first issue has been already addressed by the proposal of the nonparametric classification method and its parametrization under the restricted deterministic input, noisy, "and" gate (R-DINA) model. In this presentation, we address the second issue by proposing an adaptive implementation of the R-DINA model under the cognitive diagnostic computerized adaptive testing (CD-CAT) framework. A simulation study was conducted to compare the performance of the CD-CAT implementation of the R-DINA model with the more traditional DINA model under a range of conditions including different levels of calibration sample size, number of attributes, and item quality. We compared the DINA and R-DINA performance in terms of classification accuracy and relative measurement precision. The results indicate that the CD-CAT implementation of the R-DINA model is a promising tool to evaluate the strengths and weaknesses of students in realistic educational scenarios, offering accurate attribute classifications and estimated reliability measures in a timely fashion.

Title

An aplication of Bayesian spatio-temporal regression modeling to suicide-related 112 calls in Spain

Author(s)

Miriam Marco[1], Pablo Escobar-Hernández[1], Antonio López-Quílez[1], Francisco Sánchez-Sáez[2], María Montagud-Andrés[1], Marisol Lila[1], Enrique Gracia[1]

[1]University of Valencia, Valencia, Spain; [2]Universidad Internacional de La Rioja, Logroño, Spain

Abstract

There is a growing interest in the application of Bayesian methods to study the spatio-temporal distribution of social problems. However, the research mainly comes from Anglo-Saxon countries and others are underrepresented in the scientific literature. The aim of this study is to show an application of the Bayesian spatio-temporal modeling to suicide in an area of southern Europe Data of suicide-related 112 calls from the Valencian Community (an area of up to 5 million inhabitants of Spain) were used. Municipalities were considered as unit area (N = 542), and data from 2017 to 2022 were assessed (N = 49,792). As community-level control variables, we collected different municipal characteristics from census data: mean income, population ageing, immigration rates, residential instability, and rurality. A spatio-temporal Bayesian hierarchical Poisson regression modeling was performed using an autoregressive approach. The model incorporates two spatial random effects (the spatial autocorrelation and heterogeneity), a spatio-temporal effect and the temporal correlation. R and WinBUGS were used to run the models. Results showed a non-random spatial distribution of suicide-related calls, with areas of above-average incidence needing further exploration. In addition, results of temporal trends showed an increasing relative risk of suicide-related calls. This study illustrates a spatio-temporal epidemiological approach to study the geographical patterns and temporal trends of suicide-related 112 calls. This model can be applied to design community-level prevention strategies, as well as to guide the action protocols of the emergency resources involved in cases of suicide, such as the police or the fire brigade.

Title

Reproducibility of individual effect sizes in meta-analyses on the effectiveness of psychological interventions

Author(s)

Laura Badenes-Ribera[1], María Rubio-Aparicio[2], Julio Sánchez-Meca[2]

[1]University of Valencia, Valencia, Spain; [2]University of Murcia, Murcia, Spain

Abstract

In the last 10 years, the reproducibility and replicability of meta-analyses has become a very relevant line of research given the important role played by this type of research in the evidence-based practice approach. The large number of decisions that the meta-analyst has to make during the data extraction process to obtain or calculate effect sizes makes this stage very sensitive to errors. These errors are even more frequent in meta-analyses on the efficacy of interventions since it is necessary to extract statistical data from the primary studies to calculate effect sizes.
The purpose of this work was to reproduce the effect sizes of primary studies reported in this type of meta-analyses.
To do this, several meta-analyses were randomly selected from among 100 previously identified meta-analyses on the efficacy of psychological interventions, and statistical data from individual studies from these meta-analyses were extracted to compute effect sizes. Then, comparisons between effect sizes computed and those reported in the original meta-analyses were made.
Findings showed discrepancies between effects sizes of primary studies reported in the meta-analyses and effect sizes computed from individual studies included in these meta-analyses.
Funding. Project financed by the Region of Murcia (Spain) through the Regional Program for the Promotion of Scientific and Technical Research of Excellence (Action Plan 2022) of the Seneca Foundation - Science and Technology Agency of the Region of Murcia (grant no. 22064/PI/22). Grant PID2019-104080GB-I00 funded by MCIN/AEI/ 10.13039/501100011033.

Title

Reproducibility of meta-analytical results in psychological intervention meta-analyses

Author(s)

María Rubio-Aparicio[1], Laura Badenes-Ribera[2], Julio Sánchez-Meca[1]

[1]University of Murcia, Murcia, Spain; [2]University of Valencia, Valencia, Spain

Abstract

The purpose of this study was to assess empirically the reproducibility of the meta-analytic results of several published meta-analysis on the effectiveness clinical psychological interventions.
From a dataset consisted by 100 previously identified meta-analyses on the effectiveness of clinical psychological treatment, several meta-analyses were random selected. Primary studies included in these meta-analyses were identified and relevant data were extracted. Then, the method section of these meta-analytic studies was examined to reproduce theirs results following the same statistical methods than the original meta-analyses. Finally, comparisons between meta-analytic results reported in the original meta-analyses and the those computed from relevant data extracted from primary studies were made.
Results revealed discrepancies between meta-analytic results reported in the original meta-analyses and meta-analytic results calculated from relevant data extracted from primary studies.
Funding. Project financed by the Region of Murcia (Spain) through the Regional Program for the Promotion of Scientific and Technical Research of Excellence (Action Plan 2022) of the Seneca Foundation - Science and Technology Agency of the Region of Murcia (grant no. 22064/PI/22). Grant PID2019-104080GB-I00 funded by MCIN/AEI/ 10.13039/501100011033.

Title

Spanish Adaptation of the Personal Need for Structure Scale

Author(s)

Rafael Gil Ortega, Pablo Nájera Álvarez, Miguel Ángel Sorrel Luján, Francisco Javier Horcajo Rosado

Universidad Autónoma de Madrid, Madrid, Spain

Abstract

Introduction: The personal need for structure scale (PNS) was developed to measure individuals' desire for order, clarity, and predictability in their environment. To evaluate PNS, Thompson et al. (1989) proposed a test that has allowed researchers to investigate this construct and its relationship with other important phenomena in social psychology. Subsequently, Neuberg and Newsom (1993) created a second version by removing item 5 and creating two dimensions. In the present study, our aim was to adapt the scale and test its psychometric properties in a Spanish population sample. Additionally, we aimed to solve the debate surrounding the inclusion of item 5 and the number of dimensions that compose the scale.
Method: The PNS test was translated into Spanish by two bilingual individuals, and a back-translation process was conducted to verify its accuracy. The study sample included 735 participants, ranging from 18 to 87 years (M = 34.29, SD = 15.64). The gender distribution was 271 males, 457 females, and 7 participants with another gender identity. Several dimensionality assessment procedures were implemented to determine the number of factors underlying this psychological construct. We compared one-dimension and two-dimension structures (desire for structure and reaction to the lack of structure) and conducted a factorial analysis, test-retest reliability, convergent and divergent validity, and invariance analysis to test the Spanish version's reliability and validity.
Results: The PNS scale showed good results in terms of internal consistency and temporal stability. The results are discussed in relation to item 5 and the internal structure of the scale.

Title

Application of Artificial Intelligence in the identification of Juvenile Justice Recidivism

Author(s)

Juan García-García[1], Elena Ortega-Campos[1], Antonio Zarauz-Moreno[2], Leticia De la Fuente-Sánchez[1], Mery Estefanía Buestán-Játiva[1], María Dolores Roldán-Tapia[3], Flor Zaldívar-Basurto[1]

[1]Health Research Center (CEINSA). Dept. of Psychology. University of Almeria., Almería, Spain; [2]Dept. of Mathematics. University of Almería, Almería, Spain; [3]Dept. of Psychology. University of Almeria., Almería, Spain

Abstract

The prediction of juvenile recidivism as a research topic has increased the number and amount of resources involved in improving the response and accuracy it offers. Researchers and staff working with young offenders are interested in knowing what risk and protective factors are present in the group of youth who become recidivist offenders versus the group who do not reoffend. In the last two decades, tools have been developed to predict and manage the risk of recidivism in young offenders, focusing on the risk and protective factors present in each young person, among which YLS-CMI and SAVRY stand out. The use of new predictive models based on Artificial Intelligence is acquiring an incipient but relevant role in the field of Juvenile Justice. In this research, an Artificial Intelligence based prediction model of the factors most related to juvenile recidivism is presented. The sample is composed of juvenile offenders from a Juvenile Court in Andalusia (Spain). The objective of this work is focused on the identification of risk and protection factors for the creation of an automatic prediction system of recidivist behavior, with the aim of enabling the creation of intervention and monitoring itineraries, adapted to the needs of each young offender. The prediction models based on Artificial Intelligence open new ways of intervention of the recidivism of antisocial behavior punished in Juvenile Justice.
Project ref. P18-RT-1469, financed by the Ministry of Economic Transformation, Industry, Knowledge and Universities of the Junta de Andalucía (Spain) and FEDER funds from the European Union

Title

A Metasummary about Emotion Regulation and Cancer

Author(s)

Jennifer Pérez-Sánchez

Universidad de Salamanca, Salamanca, Spain

Abstract

Introduction: In a metasynthesis it is possible to use different strategies (data-near or data-far) to interpret the findings. The metasummary is an aggregation method (data-near) oriented towards the synthesis of qualitative research findings. The delineation of key results is relevant in order to solve ambiguities about the meaning that participants give about their own experience as well as to answer various questions of psychological interest. Discovering what strategies cancer patients use to regulate their emotions can be really useful in clinical contexts such as Psycho-oncology. Objective: The methodological aim of this work is to describe how to arrive to the main patterns of themes drawn from oncological patients' experiences. Methods: The search was carried out in Scopus, Web of Science, APA PsycINFO, APA PsycArticles, PSICODOC and Psychology and Behavioral Sciences Collection, and included the following words: "emotion regulation", "cancer", "content analysis", "thematic analysis", "framework analysis", "phenomenology", and, "grounded theory". Findings: Thirteen qualitative studies about emotion regulation and cancer published since the beginning of the 21st century were selected. Sixty themes were found out from the thirteen qualitative studies. Only ten of the themes' labels included the term "emotion". Avoidance and seeking information were two common themes. Nevertheless, they often appear in other contexts which would require a data-far strategy. Conclusions: Metasummary can be a used as a first stage towards another type of metasynthesis carried out with data-far strategies, such as reflexive thematic analysis.

Title

Difficulties in Emotion Regulation Scale: a Rasch Analysis

Author(s)

Jennifer Pérez-Sánchez, Gerardo Prieto, Ana R. Delgado

Universidad de Salamanca, Salamanca, Spain

Abstract

Research on Emotion Regulation (ER) has risen over the past two decades within the field of psychology. The Difficulties in Emotion Regulation Scale (DERS) is the most commonly used instrument to assess ER difficulties. The Spanish version of the DERS is a 28-item self-report questionnaire measuring difficulties in several abilities to regulate emotions that consists of five subscales. The format is 5-Likert response. The sample of this study was composed by 318 males aged between 20 and 69 (M age = 41.6 years, SD = 11.0). Since most studies have scored DERS by using procedures of classical test theory, in the present study the DERS answers were analysed using the Rasch Rating Scale Model (RRSM) considering that people and items can be measured in one basic dimension. Results indicated that one of the thresholds was disordered. Thus, collapsing the 5 original categories into 3 new categories was necessary. Besides, to meet the unidimensionality requirement, both the first DERS item and the awareness subscale were removed. Model-data fit was then good enough. Item Separation Reliability (ISR = 0.97) was excellent and Person Separation Reliability (PSR = 0.89) was quite good. Even though requirements of invariant measurement were met, it is worth noting that if the response categories do not perform adequately, the empirical and theoretical validity of results can be jeopardised. Moreover, it does not seem adequate to use a total score in the original version of the test when there is some evidence of a second dimension.

Title

Unifying Evidence on Delay Discounting: Open Task, Analysis Tutorial, and Normative Data

Author(s)

Sara Garofalo, Luigi Degni, Manuela Sellitto, Davide Braghittoni, Francesca Starita, Sara Giovagnoli, Giuseppe di Pellegrino, Mariagrazia Benassi

Department of Psychology, University of Bologna, Cesena, Italy

Abstract

Despite the widespread use of the delay discounting task in clinical and non-clinical contexts, several task versions are available in the literature, making it hard to compare results across studies. Moreover, normative data are not available to evaluate individual performances. The present study aims to propose a unified version of the delay discounting task based on monetary rewards and it provides normative values built on an Italian sample of 357 healthy participants. The most used parameters in the literature to assess the delay discount rate were compared to find the most valid index to discriminate between normative data and a clinical population who typically present impulsivity issues, i.e., patients with a lesion to the medial orbitofrontal cortex (mOFC). In line with our hypothesis, mOFC patients showed higher delay discounting scores than the normative sample and the normative group. Based on this evidence, we propose that the task and indexes here provided can be used to identify extremely high (above the 90th percentile for hyperbolic k or below the 10th percentile for AUC) or low (below the 10th percentile for hyperbolic k or above the 90th percentile for AUC) delay discounting performances. The complete dataset, the R code used to perform all analyses, a free and modifiable version of the delay discounting task, as well as the R code that can be used to extract all indexes from such tasks and compare subjective performances with the normative data here presented are available as online materials.

Title

Unravelling the influence of reward-associated cues on decision-making: a meta-analysis examining modulatory factors

Author(s)

Sara Garofalo, Marco Badioli, Luigi Degni, Daniela Dalbagno, Francesca Starita, Gianluca Finotti, Giuseppe di Pellegrino, Mariagrazia Benassi

Department of Psychology, University of Bologna, Cesena, Italy

Abstract

Environmental reward-associated cues (like, brands or logos) exert a powerful influence on our daily choices. Although neutral in principle, such cues acquire a motivational value through their repeated pairing with a reinforcer (e.g., a chocolate bar), and may bias future choices, driving our reward-seeking behaviour. For example, a fast-food sign may lead us to that specific fast-food to purchase and eat a hamburger, or it may lead us toward the nearest restaurant to consume food. The increasing interest in the role of predictive stimuli in guiding adaptive and maladaptive actions led to a growing number of heterogeneous evidence that require systematization and clarification. We conducted a meta-analysis of the studies currently available in literature that investigated cue-guided choice in humans. The focus was on identifying variables that can modulate the strength and direction of the effect. Our findings suggest that cue-guided choice is a robust phenomenon across a range of paradigms, but that task-specific and individual differences can modulate the strength of the effect. The results have implications for the design of interventions to improve decision-making in clinical and non-clinical populations, as well as for understanding the neural and cognitive mechanisms underlying cue-guided choice.

Title

Correcting response biases with pairwise ipsatization of graded responses

Author(s)

Rodrigo Schames Kreitchmann[1], Francisco J. Abad[2], Diego F. Graña[2]

[1]IE University, Madrid, Spain; [2]Universidad Autónoma de Madrid, Madrid, Spain

Abstract

Response biases, such as acquiescence and socially desirability (ACQ, and SD, respectively), can affect the accuracy of response data from psychological assessments. These biases can greatly distort the assessment results and compromise their validity. Several solutions have been proposed in this matter, among which we can find psychometric modelling of response styles in graded responses, person-wise standardization, or the use of forced-choice (FC) scales. In general, the FC format has been claimed to largely mitigate the problem of response biases. First, by dispensing a graded scale, FC prevents styles related with response location (e.g., ACQ). Second, when pairing statements with similar SD, it is expected to capture a most truthful response. Nonetheless, the FC format has been criticized for its lower reliability, for providing non-normative scores, and for other practical difficulties (e.g., lower convergence rate in model calibration). In this presentation, we propose a pairwise ipsatization procedure for graded responses that allows to correct response biases. In general terms, if the graded responses to two items are equally affected by the nuisance factors (e.g., response biases), the subtraction of the two responses should remove the constant nuisance term. Consequently, the pairwise response differences will be purely a function of the substantive domains. Here, we present a new probability model for pairwise response differences, which enables the recovery of normative scores. The new procedure is illustrated using empirical data and compared with the pairwise forced-choice format in terms of reliability and validity. This study provides promising results for the proposed procedure.

Title

A genuine triadic measure for capturing stress transmission and synchronization in a family

Author(s)

Shiyao Wang, Chiara Carlier, Eva Ceulemans

KU Leuven, Leuven, Belgium

Abstract

How stress synchronizes and transmits within a family over time is an important research topic. To assess such synchronization and transmission processes, many statistical methods have been developed that boil down to computing association measures. Most of the measures are dyadic, which restricts the studies to synchronization and transmission between two family members, often mother and child. However, a family is a system, in which all members might influence each other. Hence, to get a more complete understanding of stress synchronization and transmission within a family, other family members should be taken into account, starting with the father. A few studies attempted to do this by computing dyadic associations between all three pairs (mother-father, mother-child, father-child) and combining the obtained values (e.g., Bodner et al., 2018). However, it has been argued that a genuine triadic approach would yield novel insights. Our study will therefore look into triadic association measures and demonstrate which synchronization patterns are captured by which measure. We also propose a significance testing framework that accounts for serial dependence.

Title

Deep learning based IRT with missing data

Author(s)

Karel Veldkamp, Dylan Molenaar, Raoul Grasman

Universiteit van Amsterdam, Amsterdam, Netherlands

Abstract

Deep learning-based variational inference has recently gained interest as a method of estimating item response theory (IRT) parameters. These methods amortize the latent variables using a neural network, which allows for faster computation of higher dimensional IRT models. One challenge for these networks is that there is no obvious way to deal with missing data. Marginal maximum likelihood estimation (MML) can be performed using only the observed data, but neural networks have no natural way to deal with missing values. We present four different approaches to account for missing data in deep learning based IRT. We assess the performance of each method on simulated datasets for increasing levels of missingness, and compare it to marginal maximum likelihood estimation. Finally, we test the various approaches on real data with a large proportion of missing values.

Title

Hodos: Where are we going?

Author(s)

Raúl Castañeda-Vozmediano[1], Mario Calabria[2], Paula García Royo[3], Jennifer Pérez-Sánchez[4]

[1]Universidad Francisco de Vitoria, Madrid, Spain; [2]Universidad Complutense, Madrid, Spain; [3]Vall d'Hebron Institut de Recerca (VHIR), Barcelona, Spain; [4]Universidad de Salamanca, Salamanca, Spain

Abstract

The foundation of the Spanish Association of Methodology of Behavioural Sciences (Asociación Española de Metodología de las Ciencias del Comportamiento, AEMCCO) in 1993 has promoted the dissemination and meeting of relevant methodological research studies in the context of Behavioural Sciences since then. This study aims to depict the methodological research tendency based on the AEMCCO conferences. The sample of the present study consisted of the AEMCCO conference abstract books. Data storage was conducted by sorting information about the year of the conference, the number of authors, the involved universities, the type of study presentation, the aim of the study, the psychological concept of interest, and the field of research. Data analyses were performed through an iterative process to calculate descriptive statistics to show trends in methodological research topics, as well as to classify the main research topics in Behavioral Sciences over the years. Consensus among the authors was achieved. Through this study, we urge other national and international associations to examine their research trajectories to generate a resource to reflect on the way forward for methodological research in Behavioural Sciences.

Title

Internal Structure of the Sexism Against Women Gamers Scale (SAWGS): Psychometric Properties and Measure Invariance across Spanish and English versions

Author(s)

Mariela Bustos-Ortega, Hugo Carretero-Dios, Jesús L. Megías, Mónica Romero-Sánchez

University of Granada, Granada, Spain

Abstract

While gender distribution is approaching a 50:50 ratio in the global gaming community, video game culture is still dominated by masculine ideology, potentially providing breeding grounds for sexism and harassment against women online gamers. However, until now, assessment instruments focused on the construct "sexism against women gamers" are lacking. The Sexism Against Women Gamers Scale (SAWGS) is an 8-item self-report scale recently developed to take specific characteristics of in-game interactions into account when measuring sexism in online gaming. We studied the reliability, measurement invariance across gender and country, and validity of the scores of Spanish and English versions across five independent samples (N = 2,437), with participants from Spain and the United States. The SAWGS maintained a stable structure and high internal consistency reliability of scores (α's between .78 and .86) in the five samples used. Exploratory and confirmatory factor analyses showed an excellent model fit, thus confirming the one-dimensionality of the instrument. Finally, the analysis established configural, metric, and scalar invariance across gender. SAWGS was invariant at the configural level across countries (Spain and United States).

Title

Measurement invariance in factor analytic and item response theory framework

Author(s)

Patrícia Martinková[1,2], Ján Pavlech[1,2]

[1]Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic; [2]Charles University, Prague, Czech Republic

Abstract

The relations between binary factor analysis (FA) model and 2-parameter item response theory (IRT) model were established some time ago (Takane & de Leeuw, 1987; Kamata & Bauer, 2008). However, when accounting for multiple groups and testing for between-group differences on an item level, the methods based on FA and IRT frameworks are usually considered supplementary rather than equivalent. This work focuses on relations between FA models used for testing measurement invariance and IRT models used for testing differential item functioning (DIF). We focus on equivalence of the two types of group-specific models under a set of conditions. Different types of invariance are discussed, namely configural, weak (metric), strong (scalar), and partial, together with models testing for non-uniform DIF, uniform DIF, and no DIF in all items, respectively. Computational aspects and extensions to multiple groups and ordinal items are considered. A real data example is provided to demonstrate the presented relations.

Title

Collecting more data than originally planned in group sequential designs

Author(s)

Lara Vankelecom, Tom Loeys, Beatrijs Moerkerke

Ghent University, Ghent, Belgium

Abstract

Group sequential designs (GSD) are known for their efficiency in saving resources such as time and money, due to the possibility of rejecting the null hypothesis or stopping the study for futility early at an interim look. Instead of conducting a single analysis at the end of the study (like in the traditional fixed design), the data is analyzed at different time points during data collection, while controlling error rates. One flexible approach of GSD is the alpha-spending function approach, where the total number and the timing of the interim analyses need not be specified beforehand. As the amount of alpha-correction at each interim analysis depends on the fraction of information available at that point, this approach does require that the maximum number of observations is determined beforehand (for example, by performing an a-priori sample size calculation). If the final sample size of the study exceeds the predetermined maximum (i.e. overrunning), an additional (not well-known) correction of the final alpha-level is necessary. However, this correction only ensures a type 1 error rate at the nominal alpha-level when the overrunning occurs randomly (e.g. due to logistical reasons outside the researcher's control), but not when the decision for overrunning is made data-driven (e.g. when results are near significance). With a simulation study, we illustrate what the consequences are on the type I error rate of a study when 1) overrunning occurs without the additional correction and 2) when a data-driven decision strategy is used for overrunning.

Title

Early autism prediction in infants at elevated likelihood using machine learning

Author(s)

Steven Wallaert[1], Ellen Deschepper[2], Ilse Noens[3,4], Herbert Roeyers[5], Petra Warreyn[5], Lotte van Esch[3,4], Dirk De Bacquer[6]

[1]Ghent University, Department of Public Health and Primary Care, Epidemiology of Chronic Diseases Unit, Biostatistics Unit, Ghent, Belgium; [2]Ghent University, Department of Public Health and Primary Care, Biostatistics Unit, Ghent, Belgium; [3]KU Leuven, Faculty of Psychology and Educational Sciences, Parenting and Special Education Research Unit, Leuven, Belgium; [4]KU Leuven, Leuven Autism Research (LAuRes), Leuven, Belgium; [5]Ghent University, Department of Experimental Clinical and Health Psychology, Research in Developmental Disorders Lab, Ghent, Belgium; [6]Ghent University, Department of Public Health and Primary Care, Epidemiology of Chronic Diseases Unit, Ghent, Belgium

Abstract

Purpose Autism is a neurodevelopmental condition with a pervasive impact on multiple developmental domains throughout lifetime. Both children and their families could benefit from specific early support if a model would be available that reliably indicates at an early age which children are most likely to receive a diagnosis. In the TIARA study (Tracking Infants At Risk for Autism) we attempt to develop such a prediction model in infants with an elevated likelihood (EL) of developing autism.
Methods We prospectively followed 169 EL infants (101 infants having a sibling with autism, 68 preterm infants) on 5 occasions between 5 and 36 months of age, on which we collected various features (including genetic, metabolic, neurophysiological behavioral and contextual features). A best estimate research diagnosis was made, based on all available information at 36 months. Knowledge- and data-driven model building approaches are considered. Both approaches consist of building and optimizing a machine learning pipeline that is composed of missing data imputation, feature selection, dimensionality reduction, class imbalance handling, and model fitting. Bayesian optimization will be used for hyperparameter optimization. In order to estimate the performance and stability, we will apply a repeated nested cross-validation scheme where the inner loop serves to cross-validate the optimization and the outer loop serves to cross-validate the complete model building procedure.
Results We will present preliminary findings, including estimated AUC, precision, and recall of the different models.
Conclusion It is possible to apply machine learning methodology to predict autism in EL infants.

Title

Ethical artificial intelligence: proposal of a research design for the telerehabilitation of neurodevelopmental disorders

Author(s)

Aurora Castellani[1], Mariagrazia Benassi[2], Giulia Balboni[1]

[1]University of Perugia, Perugia, Italy; [2]University of Bologna, Bologna, Italy

Abstract

Neurodevelopmental disorders (NDDs) manifest early and are characterized by deficits in personal, social, and academic functioning (APA, 2013). Artificial Intelligence (AI)-based telerehabilitation techniques for NDDs have been recently developed. These interventions have several advantages, such as personalization and self-adaptivity, but also have ethical implications, such as those related to privacy or explainability (Unicef, 2020). Nevertheless, it is still unclear how ethical principles may be applied to these new interventions. This contribution proposes a research design that considers the ethical principles for using AI in the telerehabilitation of NDDs. First, a systematic review of AI-related ethical guidelines, with children being the primary users, was run to identify a valuable framework for understanding which ethical principles must be considered with children. Then, a research design for AI applied to NDDs was developed that integrated the systematic review results with the three ethical levels that should structure the AI design: a) Ethics by design. Is AI capable of ethical reasoning? b) Ethics in design. Do design methods include an assessment of the ethical implications of AI systems? c) Ethics for design. Are there standards of conduct that ensure the integrity of developers and users in all stages of system construction? (Dignum, 2018). Finally, the research design was adapted to the characteristics of children with NDDs and the telerehabilitation context. This research design may guide researchers in developing studies that consider the rules and limitations of AI for children with special needs in the telerehabilitation environment.

Title

Development and validation of an instrument for assessing parental competencies in Spanish-speaking families

Author(s)

Milagrosa Sánchez-Martín[1], Lucía Jiménez[2], Bárbara Lorence[2], Ester Herrera-Collado[2], Victoria Hidalgo[2]

[1]Universidad Loyola Andalucía, Dos Hermanas, Spain; [2]Universidad de Sevilla, Sevilla, Spain

Abstract

Parental competencies are a vital aspect of positive parenting programs, and there are several self-administered scales in the Spanish-speaking context that evaluate specific parental competencies. However, a comprehensive and ecological assessment tool that underwent rigorous design, and validation, and used a systematized administration system to collect data from families in vulnerable situations was lacking. In this study, we present the process followed to design, develop and validate an instrument for assessing parental competencies - the Parental Competence Assessment Interview (ECP-12). The first stage included focus groups with families and family intervention professionals, a systematic literature review, expert judgment, inter-observer reliability analysis, pilot testing, and exploratory analysis of dimensionality. In the second stage, the instrument was refined, and confirmed by 53 indicators assessing 12 parental competencies. We interviewed 593 parental figures (85% women; Mage = 42.19, SDage = 7.83) in charge of families at risk of psychosocial vulnerability who were users of family preservation services in Spain. Results showed that the ESEM bifactorial model was the most parsimonious Confirmatory Factor Analyses solution, considering fit indices, reliability, and various complementary indices. Additionally, evidence of validity was obtained based on the relationship with other variables, such as the Strengths and Difficulties Questionnaire (SDQ). The study confirmed that the ECP-12 is a well-defined and reliable tool for assessing parental competencies in Spanish-speaking families, with all specific competencies contributing to the parental competency construct. However, future studies are needed to refine the measurement of some specific competencies proposed in the ECP-12 independently.

Title

A comprehensive comparison of measures for assessing profile similarity at specific time points

Author(s)

Chiara Carlier[1], Julian Karch[2], Peter Kuppens[1], Eva Ceulemans[1]

[1]KU Leuven, Leuven, Belgium; [2]Leiden University, Leiden, Netherlands

Abstract

Profile similarity measures are used to quantify the similarity of two multivariate score profiles. Over half a century, computing profile similarity has increased in popularity to study for instance how two persons are similar in terms of their personality or profile of experienced emotions (over time). On the one hand, this popularity has brought many new measures into focus, yet on the other hand, many researchers stick to the known correlation and distance of scores. In this study, we have taken four steps to create a comprehensive list of measures that are useful to compare profiles and to identify meaningful groups of measures that produce similar values. During this presentation, we will focus on the last three steps. First, we have reviewed a large set of 87 similarity measures by applying them to both cross-sectional and ESM data sets and retained 43 useful profile similarity measures. Second, we have clustered these 43 measures into similarly behaving groups, and found one cluster with difference measures, one cluster with product measures and one residual cluster. Third, we have interpreted what unifies these groups and their subgroups based on theory and formulas, and linked them to concepts such as shape, scatter and elevation similarity. Last, based on these findings, we discuss some recommendations and conclusions to be drawn from this comparison with respect to the choice of measure, the Pearson correlation and centering.

Title

Causal inference in health disparities research and the "No-Multiple-Versions-of-Treatment" assumption.

Author(s)

Lizbeth Burgos Ochoa

Tilburg University, Tilburg, Netherlands

Abstract

Health disparities research is often interested in identifying the drivers of observed health differences across population groups. Such research questions are of causal nature as they involve the estimation of the effect of a certain exposure (e.g., neighbourhood poverty) on a health outcome of interest. Causal inference in health disparities research is challenging as in most scenarios researchers can only make use of observational data for this purpose. The potential outcomes framework has provided a conceptual framework and supported methodological approaches to estimate causal effects from observational data. To identify a causal effect, the framework relies on the following assumptions: exchangeability, positivity, and no-multiple-versions-of-treatment. While a large part of the literature has focused on the exchangeability assumption (and to a lesser extent on positivity), the last assumption has received little attention and it has often been taken for granted. This work focuses on the implications of the no-multiple-versions-of-treatment assumption for bias in health disparities research from a conceptual and model specification perspective. Various scenarios in which this assumption may be violated are discussed, including those related to the application of particular analytical strategies. Moreover, guidance is provided on the situations under which, even when there is a violation of the assumption, the estimates could still be interpreted as causal.

Title

HADS-Anxiety: Investigation of the scale longitudinal measurement invariance in melanoma and breast cancer patients

Author(s)

Yseulys Dubuy, Myriam Blanchin, Véronique Sébille

Nantes Université, Université de Tours, INSERM, MethodS in Patient-centered outcomes and HEalth Research, SPHERE, Nantes, France

Abstract

Background: Patient-reported outcome measures (PROM) data are challenging to analyze and interpret in longitudinal settings. Indeed, patients may give different answers to a questionnaire over time, not only because their level of the target construct has changed but also because their interpretation of the items aiming at measuring the construct has changed. For instance, cancer treatment may trigger changes in the patients' internal standard of measurement (i.e., recalibration), resulting in a lack of measurement invariance over time. In addition, those changes can differ according to the cancer site. This phenomenon is crucial to investigate as it can help to better understand psychological changes after a cancer diagnosis and to make unbiased comparisons.
Methods: We aimed to investigate longitudinal measurement invariance of the HADS-A (a scale measuring anxiety disorders) in breast cancer and melanoma patients over the year following the cancer diagnosis (ELCCA cohort NCT02893774 comprising 337 breast cancer and 110 melanoma patients). Lack of measurement invariance among the items of the HADS-A was investigated using the ROSALI algorithm (a method based on Rasch Measurement Theory to detect recalibration at the item level).
Results: Recalibration was found for one item : 'I get a (...) frightened feeling as if something awful is about to happen'; despite equal levels of anxiety disorder over time, melanoma and breast cancer patients were less likely to endorse the item one year post-diagnosis than in the month following diagnosis.
Discussion: Investigating recalibration in PROMs is crucial to better understand the psychological changes that patients experiment throughout treatments.

Title

Reinvestigating the Dimensionality of the Smartphone Addiction Scale-Short Version (SAS-SV): An Exploratory Structural Equation Modeling Approach

Author(s)

Giusy Danila Valenti[1], Palmira Faraci[2]

[1]Department of Psychological, Educational Science and Human Movement, University of Palermo, Palermo, Italy; [2]Faculty of Human and Social Sciences, University of Enna "Kore", Enna, Italy

Abstract

Although the one-dimensionality of the Smartphone Addiction Scale-Short Version (SAS-SV) is prevalent, a multidimensional structure of the scale may be reasonable because items included in the SAS-SV loaded on three separate factors (i.e., daily life disturbance, withdrawal, and overuse) on the SAS original long form. We used Confirmatory Factor Analysis (CFA) and Exploratory Structural Equation Modeling (ESEM) to examine the internal structure of the questionnaire in a sample of 484 Italian adults (51.4% males; Mage = 31.67, SD = 10.87). First, our results showed that multidimensional models outperformed the unidimensional model. Second, when compared to the 3-factor CFA, the 3-factor ESEM model [2 = 38.432; df = 18; CFI = .992; TLI = .980; RMSEA = .048 (.027-.070); SRMR = .014; AIC = 15,031.073; BIC = 15,227.631; aBIC = 15,078.457] reported lower factor intercorrelations, indicating a higher level of discrimination between factors than its competing CFA model. The occurrence of small to medium cross-loadings justified the application of the ESEM methodology. The estimation of Composite Reliability coefficients (.78 < CR < .85) and McDonald's omega (.78 < < .86) indicated good levels of internal consistency. Our findings suggest that the SAS-SV should be intended as a multidimensional measure, and they highlight the potentiality, usefulness, and appropriateness of the ESEM approach in describing and assessing psychological constructs.

Title

A new dynamic procedure for estimating item discrimination

Author(s)

Andrés González[1,2], Álvaro Postigo[1,3], Celia Serrano-Montilla[4], Luis M. Lozano[1,2]

[1]University of Granada, Granada, Spain; [2]CIMCYC, Granada, Spain; [3]University of Oviedo, Oviedo, Spain; [4]UNED, Madrid, Spain

Abstract

The item analysis is one of the essential aspects in tests and questionnaires development. Therefore, it is usual to analyze the items' discriminative power and difficulty. Focusing on the former, there are many procedures proposed to estimate it. These can be grouped into two broad categories: those based on comparing responses from extreme groups and those based on the association between the score on the item and the test. Concerning those of the first type, the effect of the item's difficulty on the maximum value of discrimination obtained is known. This effect leads to the recommendation to select items with medium difficulties to maximize their discrimination. However, because of the use of the questionnaire, it is necessary to have extremely easy or difficult items.
This research proposes a new way of estimating the discriminative capacity of items by comparing extreme groups. The key of this procedure is using a dynamic rule to conform the extreme groups based on the item difficulty. For example, in a tough item, the lower and upper groups will be formed respectively by 95% of the people with the worst performance in the test and the 5% with the best results. The analysis of this new procedure shows that items with a high capacity to discriminate at the extremes of the continuum of skill but that obtain poor results in the classic statistics are correctly identified as having good discrimination. Lastly, a generalization is proposed for the case of polytomous items.

Title

Broader autism phenotype, parental sense of competence and stress in couples with autistic children: An actor–partner interdependence moderation model

Author(s)

Patricia Recio[1], Cristina García-López[2], Pilar Pozo[1], Encarnación Sarriá[1]

[1]UNED, Madrid, Spain; [2]Hospital Sant Joan de Deu, Barcelona, Spain

Abstract

Parenting a child with autism spectrum disorder (ASD) is a demanding experience, which has been associated with heightened levels of psychological distress. Couples affect each other behaviorally but also cognitively and emotionally. The broader autism phenotype (BAQ) refers to a milder manifestation of the defining symptoms of ASD in individuals without autism. The BAQ sub-diagnostic autistic traits involve difficulties in interpersonal relationships and in pragmatic language use and rigidity traits, and it is more common in parents of individuals with autism than in the general population.
The goal of this study was to examine the relationship between actor (one's own) and partner (their partner's) parental sense of competence and their parental stress, moderated by broader autism phenotype, using actor–partner interdependence moderation model (APIM moderation).
A total of 152 mothers and fathers (76 couples) who had a child with diagnosis of ASD (aged 3 –17 years) completed assessments of broad autism phenotype, parental sense of competence, and parental stress. The goodness-of-fit for APIM moderation, using structural equation modeling, was good (chi-square(13) = 5.728, p = .956, CFI = 1.00, NFI = .956, RMSEA = 0.00, SRMR = .044).
The effect of parental sense of competence on parental stress occurred through both actor and partner pathways. The parent broader autism phenotype presented a moderation partner effect. Findings highlight the importance of considering the dyadic interdependence between parents of children with ASD toward better comprehension of parental adaptation.

Title

Psychometric properties of the State Self-Esteem Scale and its brief version scale in a Spanish sample

Author(s)

Victor Ciudad-Fernández[1,2], Tamara Escrivá-Martínez[1,2], Guadalupe Molinari Conde[3], Giulia Corno[4], Rosa Baños[1,2,3]

[1]Department of Personality, Evaluation and Psychological Treatment, Faculty of Psychology, University of Valencia, Valencia, Spain; [2]Polibienestar Institute, University of Valencia, Valencia, Spain; [3]Centre of Physiopathology of Obesity and Nutrition (CIBERobn), CB06/03/0052, Instituto de Salud Carlos III, Madrid, Spain; [4]Département de Psychologie et de Psychoéducation, Université du Québec, Outaouais, Canada

Abstract

The State Self-Esteem Scale-20 (SSES) is a commonly used questionnaire to measure fluctuations in self-esteem, but its psychometric properties have not been analyzed in a Spanish sample. Additionally, a short 6-item version (SSES-6) has been recently published. This study aimed to investigate the psychometric properties of both the SSES-20 and the SSES-6 in a Spanish sample.

A total of 821 Spanish participants were recruited for this study. In the first sample, 425 participants completed only the SSES. In the second sample, 396 participants completed the SSES along with other measures, including trait self-esteem, social desirability, depression, and state anxiety.

The results of confirmatory factor analysis suggest that a bifactor model with one general factor and three sub-dimensions (performance, social, and appearance subscales) was the best fit for our data, for both the SSES-20 (CFI=.922, RMSEA=.06, SRMR=.058) and the SSES-6 (CFI=.998, RMSEA=.001, SRMR=.013). The reliability of the SSES-20 subscales ranged from moderate to excellent (.91-.75), while the reliability of the SSES-6 subscales was lower (.61-.74). In addition, both scales were positively associated with trait self-esteem and social desirability, and negatively associated with depression and state anxiety.

In conclusion, this study provides evidence for the adequate psychometric properties of both the SSES-20 and SSES-6 in a Spanish sample. These findings suggest that these scales could be used in future research studies with Spanish-speaking populations.

Title

Assessing Myths about Cyber-Sexual Violence: First phase in the development of the AMCYS Scale

Author(s)

Rocío Vizcaíno-Cuenca[1], Mónica Romero-Sánchez[2], Hugo Carretero-Dios[1]

[1]Department of Behavioural Science Methodology, University of Granada, Granada, Spain; [2]Department of Social Psychology, University of Granada, Granada, Spain

Abstract

Introduction: Cyber-sexual violence is a prevalent and harmful form of aggression committed against women, yet little attention has been paid to the predictors of its tolerance and perpetration. In this sense, several studies have highlighted that attitudes (i.e., myths) may be an important role to explain this form of violence. Thus, the aim of this study was to carry out the first phase in the development of an instrument to assess myths about cyber-sexual violence (AMCYS).
Method: Following a theoretical rationale approach (APA, AERA & NCME, 2014), we made two qualitative studies to find content areas of the construct and create a first version of Spanish AMCYS. First, we conducted a thematic analysis of social reactions to CVS reports on Twitter. Secondly, to complement the operational definition, focus groups were carried out. After, with scale specifications, a panel of experts in methodology and sexual violence checked the comprehensibility, clarity, ambiguity and relevance to the items.
Results: We analyzed 4,046 social reactions and carried out two focus groups (one with 8 men and one with 8 women), including Spanish people who used social networking platforms daily. The results suggested four content areas (minimizing, victim blaming, exonerating the perpetrator responsibility and socio-cultural factors) that alluded attitudes that serve to justify, minimize, and deny the experiences of cyber-sexual violence. A Content Validity Index (CVI) and Kappa Index provided content validity evidences.
Conclusion: The two qualitative studies and content-based validity study made it possible to delimit the first set of items of AMCYS Scale.

Title

Development and Collection of Prior Validity Evidence for an Assessment Instrument of Police Attitudes toward Intervention in Intimate Partner Violence against Women: Mixed Method Approach

Author(s)

Celia Serrano-Montilla[1], Luis-Manuel Lozano[2], Jose-Luis Padilla[2]

[1]National Distance Education University, Madrid, Spain; [2]University of Granada, Granada, Spain

Abstract

There are no instruments in the Spanish context with adequate psychometric properties and validity evidence for measuring police attitudes toward intervention in intimate partner violence against women. The objective was to introduce the recent mixed method sequential model of scale development and validation analysis (MSDVA; Zhou, 2019) to develop an assessment instrument for their evaluation, as well as to obtain the first validity evidence for the proposed use (i.e., to detect police attitudinal patterns in order to create training programs adapted to police needs). In the Study 1 (N = 8) and the Study 2 (N = 15), we based on expert judgments to obtain content-based validity evidence for both, the operational definition and the initial set of items. In Study 3 (N = 233), the initial version of the instrument was administered to analyze the psychometric properties of the responses to the items and obtain validity evidence based on the internal structure and the relationship with other variables (myths about IPVAW, empathy, and gender-based violence specialization). The exploratory factor analysis results supported a bidimensional internal structure (proactive and reactive attitudes) consistent with the proposed semantic definition and with adequate internal consistency. In general, the items showed adequate psychometrics. In addition, evidence of validity based on the relationship with other variables were obtained.

Title

A protocol for assessing tests, scales, and questionnaires (PETEYC)

Author(s)

Isabel Benítez[1,2], Elena Govorova[3,4], Elena de la Guía[3]

[1]University of Granada, Granada, Spain; [2]Mind, Brain and Behaviour Research Center (CIMCYC), Granada, Spain; [3]2E, Estudios y Evaluaciones, Oviedo, Spain; [4]University of Oviedo, Oviedo, Spain

Abstract

Psychological and educational tests are essential for professional practice in diverse contexts. However, both the utility and the accuracy of the available instruments are not always supported by scientific evidence. In addition, professionals are not always trained to evaluate the quality of the tests before use. The present study aims to propose a protocol for guiding researchers and professionals in the assessment of tests, scales, and questionnaires. The protocol (PETEYC) pursues to provide a useful tool for helping professionals to evaluate the instrument's quality and to select the best instrument to reach the intended purpose. PETEYC was created based on two sources of information. First, a literature review focused on collecting habitual criteria for evaluating instruments' quality. The review provided information about the relevant dimensions, the habitual analysis to evaluate them, and the criterion values used to consider the results as evidence supporting the intended use of the test. Secondly, a group of experts evaluated the relevance of these dimensions for different test purposes. Experts supplied comments and suggestions about procedures to address the evaluation of the dimensions and how to interpret the results. Data from both sources were integrated to generate a protocol where the most relevant issues and limitations of a test can be identified. PETEYC is proposed as a tool for professionals to learn about the tests' limitations and to plan steps for improving the instruments' quality. A guide for using PETEYC is provided as well as indications to use it for assessing instruments' properties.

3 Thursday 13 July

3.1 Keynote speaker 10h00–11h00

Title

Latent constructs and network models in personality: implications for theory and research

Author

Marco Perugini

University of Milan–Bicocca

Abstract

Latent construct models have been historically dominant in psychology, especially in personality. In the last decade, an alternative approach using network models and tools, recently defined as network psychometrics, has become increasingly popular and influential. The talk will focus on comparing the two approaches, with their pros and cons, and articulating their implications for personality theory and research. Besides clarifying their differences and especially their different implications for theory and research, special attention will be given to how and when they can be used alongside instead of as an alternative. The main message will be that the two approaches should be seen as complementary instead of mutually exclusive, provided that their different functions are appreciated.

3.2 State-of-the-art 15h00–15h30 Aud 1

Title

Reproducibility in methodological research: Modelling and improving epistemic uncertainty

Author

Felix Schönbrodt

LMU Munich

Abstract

So far, methodologists have taken a convenient role in the replication/reproducibility crisis: Commenting from the sideline, they were able to propose best-practice procedures and solutions, and to point out weaknesses in existing research designs. But is methodological research itself immune to these issues? Can methodological research be "p-hacked" to make a study look good and increase its publishability? Recent meta-scientific research shows that methodological research has problems resembling those in other disciplines. I call for more meta-scientific research, both on methodological research itself, but also to add a methodologists' perspective on the academic system and the reform movement. Towards that goal, I suggest how epistemic uncertainty about the truthfulness of a reported result (modelled as second order uncertainty) can be a fruitful framework to understand several aspects of the replication crisis and the reform movement.

3.3 State-of-the-art 15h00–15h30 Aud 2

Title

The necessity of context in modeling complex emotion dynamics

Author

Laura Bringmann

University of Groningen

Abstract

More intensive longitudinal data is becoming available, in which people, such as patients with a clinical disorder, are measured over a long time period, for example 3 times a day for several months. These measures include emotions and contextual variables, such as feelings of sadness after a stressful interaction at work. This requires more complex modelling techniques, and it is of crucial importance that the context is included. However, in current popular theories of early warning signals and (network) studies of emotions dynamics, the focus has been on dynamics among emotions and less on context variables. I will dive into these theories, showing that there is a gap between conceptualization and theory on the one hand and statistical modeling on the other. I will then show ways forward, and how we can possibly bridge this gap, for example, by using context variables, qualitative information, and focusing on validity of measurement more generally.

3.4 Parallel sessions 08h30–10h00 Auditorium 1

Symposium Overview

Structural Equation Models with Machine Learning and Data Mining: Some Recent Developments and Software Packages

Author(s)

Christoph Kiefer[1], Manuel Arnold[2]

[1]Bielefeld University, Bielefeld, Germany; [2]Humbold University Berlin, Berlin, Germany

Abstract

Structural equation modeling is one of the most popular statistical frameworks in the social and behavioral sciences. With the advent of big data, the number of potentially interesting variables one wishes to include into a structural equation model (SEM) has increased. This poses two challenges for traditional SEM methods: (a) including too many variables and parameters into a SEM can lead to an overly complex or overfitting model and (b) neglecting valuable information can lead to unobserved heterogeneity. In this symposium, we shed light on recent developments bringing machine learning and data mining techniques to structural equation modeling to address these challenges. The first talk will present how regularization strategies can help to identify relevant parameters in complex SEMs. The Julia package StructuralEquationModels.jl will be presented as a viable software implementation for regularization in SEM. The second talk will present how SEM trees can be used to assess and visualize heterogeneity in SEM. A new approach will be presented, which allows to explore heterogeneity with regard to several parameters. The third talk will present the SubgroupSEM approach, a recently proposed alternative for assessing heterogeneity in SEM. The talk will provide an introduction to SubgroupSEM, including a comparison to SEM trees and an empirical illustration. The fourth talk will present how the SubgroupSEM approach can be used to explore heterogeneity of treatment effects in non-randomized experiments. The approach allows simultaneous accounting for confounders and moderators of the treatment effects.

Title

Subgroup Discovery in Structural Equation Models

Author(s)

Axel Mayer, Christoph Kiefer

Bielefeld University, Bielefeld, Germany

Abstract

Often, the detection of groups with distinct sets of parameters in structural equation models (SEM) is of key importance for applied researchers – for example, when investigating differential item functioning for a mental ability test or examining children with exceptional educational trajectories. Two common approaches that can be used for this purpose are structural equation mixture modeling (SEMM) and structural equation model trees (SEMtrees). Both allow for discovering (potentially latent) subgroups that are distinct with regard to their structural relations. SEMM uses a latent class approach while SEMtrees is based on recursive partitioning. In this talk, we present a recently proposed alternative approach, combining subgroup discovery – a well-established toolkit of supervised learning algorithms and techniques from the field of computer science – with SEM. We provide an introduction on what distinguishes subgroup discovery from common approaches and how subgroup discovery can be applied to detect subgroups with exceptional parameter constellations in SEM based on user-defined interestingness measures. The approach is illustrated using both artifical and real-world data from a educational large-scale assessment study. The illustrative examples were conducted in the R package subgroupsem, which is a viable implementation of our approach for applied researchers.

Symposium title

Structural Equation Models with Machine Learning and Data Mining: Some Recent Developments and Software Packages

Title

Using SubgroupSEM for Finding Subgroups with Unique Treatment Effects in Non-Randomized Experiments

Author(s)

Benedikt Langenberg[1], Christoph Kiefer[1], Florian Lemmerich[2], Axel Mayer[1]

[1]Bielefeld University, Bielefeld, Germany; [2]University of Passau, Passau, Germany

Abstract

We recently proposed a new approach for efficient subgroup discovery in structural equation models called SubgroupSEM. It is an exploratory technique that helps generate hypotheses by detecting unique groups with distinct parameter sets (i.e., distinct combinations of covariates). For instance, when comparing a treatment to a control arm in a non-randomized experiment, SubgroupSEM can be used to identify groups that particularly benefited from the treatment. The effects in the detected groups, however, cannot be interpreted causally. That is, subgroup discovery algorithms are prone to erroneously finding or overlooking unique subgroups when covariates are associated with unidentified confounders that affect both the group assignment and the dependent variable.
To address this issue, we extend the SubgroupSEM approach to include propensity scores, which help account for confounders. In this presentation, we illustrate the use of SubgroupSEM with propensity scores through a hypothetical example involving individuals who received or did not receive psychotherapy after a cancer diagnosis. We compare depressive symptoms between the two groups and use SubgroupSEM to identify groups that particularly benefited from the treatment. We present scenarios that include confounders where subgroup discovery algorithms either identify non-unique subgroups or overlook unique subgroups. Additionally, we compare results from SubgroupSEM with and without propensity scores and discuss the implications for covariate selection and causal inference.

Symposium title

Structural Equation Models with Machine Learning and Data Mining: Some Recent Developments and Software Packages

Title

Regularized Structural Equation Modeling with StructuralEquationModels.jl

Author(s)

Maximilian S. Ernst[1], Aaron Peikert[1,2,3], Andreas M. Brandmaier[1,3,4]

[1]Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany; [2]Humboldt-Universität zu Berlin, Department of Psychology, Berlin, Germany; [3]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, London, United Kingdom; [4]Department of Psychology, MSB Medical School Berlin, Berlin, Germany

Abstract

Researchers using structural equation models (SEM) often find themselves in scenarios where setting parameters to zero or estimating them freely is not decidable from theory alone. As a result, they end up with more complex models than their sample size allows. Regularization alleviates this problem by adaptively shrinking unneeded parameters towards zero. This allows fitting models with many parameters to limited data, using a data-driven approach to impose sparsity. In recent years, regularized SEM have gained popularity. For example, they were proposed as an alternative for exploratory factor analysis or to reduce the error of parameter estimates in settings with high multicollinearity.
However, efficient and reliable software implementations are still lacking. For this purpose, we present StructuralEquationModels.jl, a software package written in the Julia language that provides a flexible and efficient implementation of regularized SEM. It makes many different forms of regularization (e.g., l1, l2, elastic net, l0, infinity norm) available, as well as the possibility to extend the package for new forms of regularization easily. In addition, it is orders of magnitude faster than previous implementations of regularized SEM.

Symposium title

Structural Equation Models with Machine Learning and Data Mining: Some Recent Developments and Software Packages

3.5 Parallel sessions 08h30–10h00 Auditorium 2

Symposium Overview

New frontiers in neuropsychological assessment

Author(s)

Pasquale Anselmi, Debora de Chiusole

University of Padova, Padova, Italy

Abstract

The symposium presents advanced methods and procedures for the assessment of neuropsychological functions, including executive functions and fluid intelligence. It explores the advantages offered by knowledge space theory, procedural knowledge space theory, and item response theory to develop and administer instruments for the accurate and efficient assessment of an individual's capabilities and functioning. In the case of item response theory, the assessment results in a number that expresses the trait level of the individual. In the case of knowledge space theory and procedural knowledge space theory, the assessment results in a collection that identifies the problems that the individual is capable of solving or the skills that the individual has available. Moreover, the symposium presents the first results obtained within a funded Italian research project where knowledge space theory and procedural knowledge space are used to develop and adaptively administer new web-based tools for the assessment of executive functions and fluid intelligence, which are based on the Tower of London and the Raven matrices, respectively.

Title

Usefulness of item response theory in the choice and development of neuropsychological tests

Author(s)

Pasquale Anselmi[1], Alice Bacherini[2], Giulia Balboni[2], Egidio Robusto[1]

[1]University di Padova, Padova, Italy; [2]University of Perugia, Perugia, Italy

Abstract

Different neuropsychological tests might be optimal for different purposes. Some tests are especially suited to reliably measuring the latent trait levels of individuals, whatever they are. Other tests are especially suited to reliably classifying individuals into one of two groups (e.g., impaired vs. nonimpaired). Having clear the intended use of the test is thus essential when choosing a test among a series of alternatives or when developing it. The present talk illustrates the usefulness of item response theory (IRT) in the choice and development of tests. Unlike classical test theory, IRT focuses on single items and assumes that the precision with which a latent trait is measured varies across the trait levels. Two relevant concepts within the theory are item information function (IIF) and test information function (TIF). The IIF shows how well and precisely a particular item measures the latent trait at various trait levels. By aggregating the IIFs across all items of the test, the TIF shows us how well and precisely the test measures the latent trait at various trait levels. If the test should be used for reliably measuring the latent trait levels of individuals, then the IIFs should be more or less evenly spread over the most crucial part of the trait range. If the test should be used for reliably classifying individuals, the IIFs should be concentrated around the chosen cutoff. The results of a simulation study and an empirical application are presented and discussed.

Symposium title

New frontiers in neuropsychological assessment

Title

Deconstructing the Tower of London: A Systematic Analysis of the Tower of London problem space

Author(s)

Andrea Brancaccio, Luca Stefanutti, Pasquale Anselmi, Debora de Chiusole, Marina Ottavia Epifania

Department of Philosophy, Sociology, Pedagogy, and Applied Psychology University of Padua, Padova, Italy

Abstract

The Tower of London(TOL) test has been employed in several neuropsychological assessment batteries. A problem of the TOL consists of matching an initial configuration (i.e., a spatial disposition of the balls on the pegs) with a goal configuration using the minimum number of moves. The original TOL test proposed by Shallice (1982) was composed of 12 problems with the same initial state and different goal states.
Recently, Stefanutti, de Chiusole & Brancaccio (2021) proposed a new way to model the family of solutions
for all TOL problems using Procedural Knowledge Space Theory (PKST; Stefanutti, 2019). In PKST, such a family is called a problem space. Different from the original TOL version, in PKST, the goal state is fixed for all the problems whereas the initial states vary.
In this study, a systematic analysis of the problem's characteristics in a problem space (i.e., the minimum number of moves to solve the problem, the number of alternative solution paths, and the hierarchy of the initial configuration) was conducted. It is well-known in the literature that these characteristics affect the difficulty of TOL problems. This systematic analysis led to the definition of a test containing 35 problems that maximizes the information in an adaptive assessment while minimizing the number of problems asked. Finally, preliminary data about validating a probabilistic PKST model for analyzing responses to the TOL test are presented.

Symposium title

New frontiers in neuropsychological assessment

Title

Convergent and divergent validity of the new web measure of executive functions AdapTol

Author(s)

Irene Pierluigi, Alice Bacherini, Giulia Balboni

University of Perugia, Perugia, Italy

Abstract

The assessment of executive functions is fundamental for identifying strengths and weaknesses and planning early rehabilitative interventions. However, individuals with neurodevelopmental disorders (e.g., autism spectrum disorder) or psychiatric conditions (e.g., schizophrenia) frequently struggle with standard testing procedures that take a long time. Better assessment ways, adequate for these populations' specific needs, are recommended to improve the effectiveness and precision of the evaluation. A computer-based adaptive assessment might be a possible solution. Using the principles of the Knowledge Space Theory (e.g., Doignon & Falmagne, 2011), a new adaptive measure of executive functions (called AdapTol) is being developed. This contribution aims to investigate the convergent and divergent validity of AdapTol in comparison with a traditional measure of planning abilities (i.e., the Tower of London [TOL]) and a measure of fluid intelligence (i.e., Raven matrices), respectively. The three instruments are being administered to about 440 individuals aged 4-18 of the general Italian population: 60 preschoolers, 200 elementary school students, 100 middle school students, and 80 secondary school students. The data collection is still in progress. In agreement with the guidelines of the European Federation of Psychologists Association (2013), a Pearson's correlation coefficient .55 is expected among the scores of the AdapTol and those of the traditional TOL, while smaller magnitudes are expected with the Raven matrices. The magnitude of the correlation coefficients will be further compared using the Williams t-test for dependent overlapping correlation coefficients (Steiger, 1980).

Symposium title

New frontiers in neuropsychological assessment

Title

Psychometric properties of the Italian adaptation of the System Usability and Acceptance Model scale for children as users of AdapTol

Author(s)

Matilde Spinoso, Noemi Mazzoni, Matteo Orsoni, Sara Garofalo, Mariagrazia Benassi, Sara Giovagnoli

Department of Psychology "Renzo Canestrari", University of Bologna, Bologna, Italy

Abstract

The perceived ease of use and attitude towards new technologies are particularly relevant for the effectiveness and implementation of new digital neuropsychological tests. The present work aims to evaluate the psychometric properties of the Usability and Attitude Scale (UAS). The UAS is an Italian adaptation of the System Usability Scale (SUS) and Technology Adaptation Model Scale (TAM) for children, aimed at assessing the attitude and usability of AdapTol, a new computerized test for the evaluation of executive functions. 361 children participated in the study (age 4-8: n=177; age 9-13; n= 184). All participants responded to AdapTol, a new digital assessment tool based on Tower of London Test and then completed the UAS. UAS includes eight modified selected items from TAM and SUS. Exploratory Factor Analysis was performed in the two age groups (4-8 and 9-13) separately. In both analyses, the 3 factors extracted explained the 68% of the total variance. The 3 factors included those items dealing with usability (F1, 3 items), those expressing the perceived ease of use of the tool (tablet) (F2, 3 items), and those for the usability perceived by peers (F3, 2 items). The questionnaire showed adequate internal consistency (Cronbach's Alpha =. 84-.83). In conclusion, the UAS showed good content validity and internal consistency.

Symposium title

New frontiers in neuropsychological assessment

3.6 Parallel sessions 08h30–10h00 Auditorium 3

Title

Exploring Different Mixed-Effects Models for Approximating Time-Varying Experimental Effects

Author(s)

Salome Li Keintzel[1], Anna Nikolei[1,2], Florian Scharf[1]

[1]University of Kassel, Kassel, Germany; [2]University of Münster, Münster, Germany

Abstract

Experimental psychologists are often interested in experimental effects decreasing or increasing over the course of the experiment, for example due to habituation or fatigue. Typically, the exact time course of the experimental effect is unknown. Different mixed-effects models with varying degrees of flexibility could be used for approximation, ranging from simple linear mixed-effects models to low-order polynomial mixed-effects models or mixed-effects splines. Recently, an even more flexible method was proposed by combining the mixed-effects approach with non-parametric regression trees (RE-EM Tree, Hajjem et al., 2011; Sela & Simonoff, 2012). We conducted a simulation study that mimicked a realistic reaction time experiment to examine under which circumstances these models can recover non-linear temporal variation in experimental effects. Performance was compared for different monotonous and non-monotonous time courses of the experimental effect. We found that linear and low-order polynomial mixed-effects models often achieved a reasonable approximation depending on the simulated time course. Mixed-effects Splines mostly offered no substantial improvement. Remarkably, the method deemed most flexible, RE-EM Trees, often failed to recover any time course at all and when it did, the approximation remained simplistic.

Oral presentations session title:

Psychometrics

Title

A first measurement of between-school pupil mobility in the Flemish primary education market: methodological issues

Author(s)

Georges Van Landeghem

KU Leuven, Leuven, Belgium

Abstract

The intensity of between-school pupil mobility in Flemish regular primary education is measured for the first time, using enumerative data about the individual trajectories in a birth cohort. Given the lack of internationally established measures, several indicators representing the system's, the pupil's, and the school's point of view are proposed. It turns out that the level of mobility is significant, as compared to other atypical events in pupils' trajectories. The consequent importance of mobility as an often overlooked nuisance factor in education research is discussed. The under-researched issue of the ambiguity of school definitions is shown to have an impact on mobility measurement in Flanders and is discussed from the wider perspective of educational research.

Oral presentations session title:

Psychometrics

Title

A new perspective on test norming

Author(s)

Andries van der Ark

University of Amsterdam, Amsterdam, Netherlands

Abstract

In the last two decades, there have been great advancements in norming psychological tests. Regression-based norming has become the standard, and regression models have become more and more flexible, which helps to avoid violations of model assumptions and biased norms. Norming methods usually rely on regression models with a continuous response variable, defined on the real line, whereas test scores are typically discrete with a restricted range. We propose a norming method that provides a discrete and range-preserving estimate of the test-score distribution. Suppose that the test consists of J items with item scores $X_1, \ldots, X_j$. Let $X = X_1 + \ldots + X_j$ denote the test score, let $\mathbf{Z} = (Z_1, \ldots, Z_n)$ denote the predictors for which separate norms should be constructed, and let g denote a density function. We consider the task of norming a psychological test equivalent to estimating $g(X|\mathbf{Z})$. From $g(X|\mathbf{Z})$, the desired norm may be derived (e.g., percentile ranks, stanines). To achieve a discrete and range-preserving estimate we estimate the joint density $g(X_1, \ldots, X_j, \mathbf{Z})$ using a latent class model or—if $\mathbf{Z}$ contains continuous variables—a general location model, and transform $g(X_1, \ldots, X_j, \mathbf{Z})$ to $g(X|\mathbf{Z})$. For a real-data set containing the scores of the SPARTS (a test measuring teacher-student relationships) and several covariates, and for simulated data, we compare the results of the proposed norming method to regression-based norming with GAMLSS, and traditional norming

Oral presentations session title:

Psychometrics

3.7 Parallel sessions 08h30–10h00 Auditorium 4

Title

New analytic rotations for bifactor modeling and metric invariance in Exploratory Factor Analysis

Author(s)

Marcos Jiménez[1], Francisco Abad[1], Eduardo García-Garzón[2], Luis Eduardo Garrido[3], Vithor Franco[4]

[1]Universidad Autónoma de Madrid, Madrid, Spain; [2]Universidad Camilo José Cela, Madrid, Spain; [3]Pontificia Universidad Católica Madre y Maestra, Santiago De Los Caballeros, Dominican Republic; [4]Universidade São Francisco, São Paulo, Brazil

Abstract

For the last two decades, the gradient projection algorithm (GPA) has been the most used algorithm to rotate factor solutions. However, GPA is currently limited to perform either orthogonal or oblique rotations. This is a shortcoming because in some cases the researcher knows in advance that some factor correlations should be zero whereas others could be nonzero. Thereby, we developed a new kind of rotation, the partially oblique rotation, in which both oblique and orthogonal factors co-exists. This rotation is desirable to estimate exploratory bifactor models with multiple general factors in which the general factors may be correlated but should remain orthogonal to the group factors, and in exploratory multitrait-multimethod models, in which the methods and traits should be uncorrelated between them but could correlate among themselves. Additionally, we used this new rotation to create a method for estimating exploratory factor models with metric invariance between groups. Some empirical examples are presented to illustrate these new developments with the `bifactor` R package.

Oral presentations session title:

Exploratory Factor Analysis

Title

Modeling latency differences in Exploratory Factor Analyses for ERP data

Author(s)

Kim-Laura Speck, Florian Scharf

Universität Kassel, Kassel, Germany

Abstract

Event-related potentials (ERPs) represent brain activity for multiple sampling points at multiple electrodes for participants in different (experimental) groups. Exploratory factor analysis (EFA) is commonly applied to analyze ERP data and provide reliable group effect estimates. EFA assumes invariant factor loadings for all observations (measurement invariance), that is, invariant time courses in ERP components for electrodes, participants and groups. Measurement invariance is unlikely for ERP data which often comprise latency differences between participants or groups (i.e., shifts in factor loading patterns across sampling points). Ignoring latency differences during the analysis results in biased effect estimates and bears the risk of erroneous substantive conclusions. We evaluate and compare two approaches to take latency differences on the participant level into account. First, latency differences between participants can result in an additional factor with a known loading pattern that can be mathematically approximated (Möcks, 1986) to get unbiased condition effects. We suggest a customized rotation method to control for latency differences by extracting this additional factor. Second, participant-specific latencies for each factor can be explicitly modeled in an extended factor analytic approach that is called the shifted factor analysis (SFA, Hong & Harshman, 2003). We compared these two approaches regarding bias in estimated group effects and feasibility in an applied research setting.

Oral presentations session title:

Exploratory Factor Analysis

Title

An evaluation of the Nest Eigenvalue Sufficiency Test (NEST)

Author(s)

Pier-Olivier Caron

Université TÉLUQ, Montréal, Canada

Abstract

Determining the number of factors to retain in an exploratory factor analysis is an open methodological problem still after 75 years of research. A plethora of techniques exist to address this challenge. One of the most promising techniques is the Next Eigenvalue Sequence Test (NEST; Achim, 2017), which shows excellent performance (Achim, 2020; Brandenburg & Papenberg, 2022), but has not, however, been systematically compared to well-established competitors (see, Auerswald & Moshagen, 2019). The present study thus proposes a simulation with synthetic factor structures to compare NEST, parallel analysis, minimum average partial correlation, 2 sequential test, Hull method, and the empirical Kaiser criterion. The structures are based on 24 variables containing 1 to 8 factors, with loadings ranging from .40 to .80, inter-factor correlations from .00 to .30, on three sample sizes, 120, 240 and 480. In total, 360 scenarios are tested 1000 times. Performance is evaluated in terms of accuracy (correct identification of dimensionality), bias (tendency to over- or underestimate dimensionality) and variability (magnitude of prediction error). The results show that NEST outperforms these competitors. While most techniques do well in easy scenarios, NEST particularly stands out in difficult ones. Scenarios in which all methods fail are discussed. Some limitations of NEST are adressed. Finally, a new R package, named RNest, which implements NEST is promoted.

Oral presentations session title:

Exploratory Factor Analysis

Title

Exploring the measurement model in (high-dimensional) multigroup data: Regularized joint latent variable analysis

Author(s)

Katrijn Van Deun, Trà Lê

Tilburg University, Tilburg, Netherlands

Abstract

Exploring multi-group data for similarities and differences in the measurement model is a substantial part of the research conducted in the behavioral and social sciences: It is relevant to studying measurement invariance of scales and group-differences in multivariate relations (e.g., biological pathways in psychopathology). Yet, currently available methods are restrictive in their use and do not scale up with the increasing complexity of current research paradigms. On the one hand, these methods cannot handle data with small sample sizes relative to the number of variables while such high-dimensional data (e.g., thousands of biomarkers for several dozen of patients or word counts obtained by text mining of tweets) are more and more used as a result of digitalization. On the other hand, users of software for exploratory multigroup methods might encounter issues such as the need to fix some parameters beforehand, a lack of convergence, and inconclusive results. Here, we propose a regularized latent variable method that addresses these issues by building on a strong computational framework: The resulting method yields solutions that are constrained to show simple structure and similarity of the loadings over groups when supported by the data. The minimal required input by the user is restricted to the data and number of latent variables. Interpretation is eased by exact zero loadings and, when measurement invariance holds, exactly equal loadings over groups. The (comparative) performance of the method is evaluated in a simulation study and illustrated on empirical data.

Oral presentations session title:

Exploratory Factor Analysis

3.8 Parallel sessions 08h30–10h00 Lecture room 1.2

Title

Local Equating of Test Scores using Propensity Scores: A New Method for Non-Equivalent Test Groups Without Anchor Items

Author(s)

Gabriel Wallin[1], Marie Wiberg[2]

[1]London School of Economics and Political Science, London, United Kingdom; [2]Umeå University, USBE, Umeå, Sweden

Abstract

Test score equating is a family of statistical methods aiming to adjust for differences in difficulty between test forms in standardized educational testing, aiding fair comparisons of examinees. This study explores the usefulness of covariates on equating test scores when the test groups are samples from different populations and no common items (i.e., anchor items) are available. The covariates are captured by an estimated propensity score, which is used as a proxy for latent ability to balance the test groups. We propose a new method for local equating, which uses a family of equating transformations instead of a single transformation. The idea is to approximate as closely as possible the individual members of a family of true equating transformations using all empirical information in the equating design. The proposed method is illustrated in an empirical study and evaluated in a simulation study which shows several realistic scenarios where it provides accurate and reliable estimates of the equating function. However, it also sheds light on the challenge of making fair comparisons between non-equivalent test groups in the absence of common items. The study identifies scenarios where equating performance is acceptable and problematic, provides practical guidelines, and identifies areas for further investigation. Our local equating method thus provides a promising alternative to traditional methods of equating when common items are not available, and which is easy to implement. This study has important implications for test developers and practitioners who need to equate test scores across different groups.

Oral presentations session title:

Item Response Theory (IRT)

Title

How to do computer simulations that scientific journals will (probably) not like

Author(s)

Ivailo Partchev

Cito, Arnhem, Netherlands

Abstract

Although academic psychometrics and practical testing share their subject and a large part of the methods, they are distinct activities with different purposes, circumstances, and priorities. Whilst the logic of scientific inquiry inevitably places emphahis of goodness of fit, testing is in need of fair and transparent rules known in advance of the "game".
We describe the logic of a small simulation study designed from the perspective of testing and explain the motivation behind each step. The results shed light on a number of phenomena, such as the negative correlation between item difficulties and item discriminations often observed in practice, some well-known problems with estimating the 3PL model, and more.

Oral presentations session title:

Item Response Theory (IRT)

Title

Modeling item responses under different frameworks

Author(s)

Patrícia Martinková[1,2], Adéla Hladká[1]

[1]Institutue of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic; [2]Charles University, Prague, Czech Republic

Abstract

Item responses may be modeled in the factor analytic framework as well as in the framework of generalized linear and nonlinear mixed-effect models (also referred as item response theory, IRT). In this work, we first discuss the relationships between the two frameworks and advantages of each of them. We then focus on the latter one, and describe a step-by-step development of IRT models via empirical characteristic curves and generalized linear and nonlinear models (GLNM), while emphasizing the didactic value of such an approach (Martinková & Hladká, 2023). We outline possible further uses of GLNM in testing criterion-related item validity and we demonstrate them with real data examples. Finally, we present some novel approaches to parameter estimation in the GLNM framework and discuss their challenges in practical implementation.
Martinková, P., & Hladká, A. (2023) Computational Aspects of Psychometric Methods. With R. Chapman and Hall/CRC (In Press). ISBN 9780367515386.

Oral presentations session title:

Item Response Theory (IRT)

Title

A Highly Adaptive Testing Design for PISA

Author(s)

Andreas Frey, Christoph König, Aron Fink

Goethe University Frankfurt, Frankfurt, Germany

Abstract

From 2018 on the Programme for International Student Assessment (PISA) switched from using linear test forms to using multi-stage testing (MST). This transition led to a relatively small increase of 4–7% in terms of test information compared to the mode of item presentation used before. We will present an alternative highly adaptive testing (HAT) design for PISA, which follows the principle to be as adaptive as possible when selecting items while accounting for PISA's nonstatistical constraints and addressing issues concerning PISA such as item position effects. HAT combines several established methods from the area of computerized adaptive testing. The HAT-design was compared to the PISA 2018 multistage design (MST) in a simulation study based on a full factorial design with the IVs response probability (RP; .50, .62), item pool optimality (PISA 2018, optimal), and ability level (low, medium, high). PISA-specific conditions regarding sample size, missing responses, and nonstatistical constraints were implemented. HAT clearly outperformed MST regarding test information, RMSE, and constraint management across ability groups but it showed slightly weaker item exposure performance. Raising RP to .62 did not decrease test information much and is therefore a viable option to foster students' test-taking experience with HAT. Test information for HAT was up to three times higher than for MST when using a hypothetical optimal item pool. Results on the effects of HAT on population estimates using plausible values will also be presented at the conference. Summarizing, HAT proved to be a promising and applicable test design for PISA.

Oral presentations session title:

Item Response Theory (IRT)

3.9 Parallel sessions 08h30–10h00 Lecture room 1.3

Title

Evaluating robust variance estimation in MASEM

Author(s)

Zeynep Bilici, Suzanne Jak

University of Amsterdam, Amsterdam, Netherlands

Abstract

Dependent effect sizes in meta-analysis are quite common; studies may measure the same constructs across different time points, using different operationalization strategies or by using multiple informants. Whereas traditional meta-analysis can deal with these dependencies more easily, when researchers are trying to meta-analyze multiple relationships in a given SEM model the dependencies are more complex to deal with. In the context of MASEM, when we have multiple effect sizes available for the same relationship in the same study, some of the methods used in the context of traditional meta-analysis is still applicable, such as aggregation, elimination and ignoring dependency. Previous simulation results comparing the methods of aggregation, elimination, ignoring dependency and univariate three-level modeling (Wilson et al., 2016) in the context of MASEM show that there is not one method that performs well across different conditions and evaluation criteria. Robust variance estimation (RVE) suggests an alternative approach, whereby the covariances in sampling errors are estimated by averaging the cross-products of residuals within each study (Hedges et al., 2010). By integrating a SEM model in multivariate meta-analysis with robust variance estimation, we aim to assess the problem of dependent correlations in MASEM. The current simulation study assesses the performance of RVE across conditions of varying number of studies, number of dependent effect sizes within studies, the magnitude of the correlation between the dependent effect sizes and the between studies variance.

Oral presentations session title:

Meta-Analysis

Title

Can we Include Dichotomous Predictor Variables in Meta-Analytic Structural Equation Modeling?

Author(s)

Hannelies de Jonge, Kees-Jan Kan, Suzanne Jak

University of Amsterdam, Amsterdam, Netherlands

Abstract

Meta-analytic structural equation modeling (MASEM) allows a researcher to simultaneously examine multiple relations among variables by fitting a structural equation model to a meta-analytic dataset. Several relationships may be hypothesized between a predictor (X), a mediator (M), and an outcome variable (Y). Within such a model, X can be a dichotomous variable, allowing researchers to examine, for example, the direct and indirect effects of an intervention as in randomized controlled trials (RCTs). One obstacle would be that MASEM requires correlation (or covariance) matrices as input, whereas the summary statistics reported in RCTs typically concern standardized mean differences (e.g., Cohen's d or Hedges' g). For MASEM, we need to convert the standardized mean difference to a point-biserial correlation. Possible conversion formulas vary across publications, statistical software, and online conversion tools, and it is unclear which one is most appropriate for use in MASEM. We (re-)considered the Cohen's d and Hedges' gto point-biserial correlation relationships and commented on the assumptions underlying the various expressions. We will conduct a simulation study to investigate the behavior of the various conversion formulas under several conditions: The number of primary studies, the studies' sample sizes, and whether correlation matrices are complete or not. We intend to present the results of our simulation study at the conference. We plan to develop a user-friendly web application that converts the user's primary study statistics into an effect size suitable for use in MASEM.

Oral presentations session title:

Meta-Analysis

Title

Nonparametric estimation of heterogeneity in rare events meta-analysis using arm-based and contrast-based approaches

Author(s)

Katrin Jansen, Heinz Holling

University of Münster, Münster, Germany

Abstract

Modelling heterogeneity in meta-analysis of binary data is a challenging endeavor when events are rare. In this situation, conventional random effects models often fail to detect relevant heterogeneity or to provide unbiased estimates of the between-study variance. In this talk, we explore the potential of nonparametric mixture models, which provide an alternative approach to model heterogeneity in rare events meta-analysis. Specifically, we compare arm-based and contrast-based nonparametric mixture models. We show how these models can be estimated using the Expectation-Maximization algorithm, and highlight their differences. Results of a simulation study will be presented in which we compare the approaches in terms of their estimation of the pooled effect and the between-study variance. Finally, we evaluate how well model selection criteria, such as the AIC and BIC, distinguish between arm-based and contrast-based approaches.

Oral presentations session title:

Meta-Analysis

Title

Confidence Intervals for the Amount of Heterogeneity Accounted for in Meta-Regression Models

Author(s)

Wolfgang Viechtbauer

Maastricht University, Maastricht, Netherlands

Abstract

The effect sizes included in a meta-analysis often exhibit more variability than would be expected based on sampling variability alone. This suggests the presence of 'heterogeneity', that is, variability in the underlying true effects. Mixed-effects meta-regression models are then typically used to examine if the heterogeneity can be accounted for by one or multiple predictor ('moderator') variables (Thompson & Sharp, 1999). Researchers then also often compute a pseudo R2 statistic (Raudenbush, 2009) to estimate how much of the heterogeneity has been accounted for by the model. However, the estimate is often imprecise, especially when the number of studies is small (Lopez-Lopez et al., 2014). A corresponding confidence interval could be used as an indicator for the precision of this estimate. A variety of different methods for constructing such a confidence interval were examined and compared to each other by means of a Monte Carlo simulation study. The methods examined included parametric and non-parametric bootstrapping and using standard procedures from regular regression models (based on the inversion of the F-statistic). The number of studies, amount of heterogeneity, and true R2 value were systematically varied across 176 different conditions. The performance of the methods was examined in terms of their coverage rate and average interval width.

Oral presentations session title:

Meta-Analysis

3.10 Parallel sessions 11h30–13h00 Auditorium 1

Symposium Overview

Measuring changing abilities: Psychometrics for adaptive learning

Author(s)

Maria Bolsinova

Tilburg University, Tilburg, Netherlands

Abstract

Adaptive learning systems (ALS, a.k.a. adaptive learning environments, computer adaptive practice systems) are designed to dynamically adjust the level of practice and instructional material based on the individual learners' abilities in order to improve both the learning process and the learning outcomes. To optimize feedback, instructions, and learning material, one needs to have continuously updated, accurate and reliable measures of the students' changing abilities. This makes measurement one of the central issues in ALS. Measuring change is in itself not a trivial problem, however the adaptive and large-scale nature of ALS pose additional challenges for traditional psychometric models and algorithms. In this symposium we discuss these methodological challenges and present some of the solutions that have been developed in the area of psychometrics for adaptive learning in the recent years. The first presentation explains why traditional psychometrics cannot be readily adopted in ALS and shows how an extension of the Elo rating system originating from chess can be a solution. The second presentation shows that even with this methodological innovation measurement issues can arise and discusses how they can be addressed in practice. The third presentation discusses the Urning algorithm which has been developed in adaptive learning as an alternative to Elo and focuses on how the "cold-start" problem which is omnipresent in ALS can be dealt with within this algorithm. The forth presentation goes beyond measurement in itself and focuses on modeling learning efficiency with longitudinal IRT in order to evaluate the added value of adaptive learning for students.

Title

Methods to alleviate the cold-start problem of adaptive learning systems using Urnings algorithm.

Author(s)

Bence Gergely[1,2,3], Han van der Maas[4], Gunter Maris[5], Maria Bolsinova[6]

[1]Eötvös Lóránd University, Doctorate School of Psychology, Budapest, Hungary; [2]Eötvös Lóránd University, Institute of Psychology, Budapest, Hungary; [3]Károli University of the Reformed Church, Department of General Psychology and Methodology, Budapest, Hungary; [4]University of Amsterdam, Department of Psychological Methods and Statistics, Amsterdam, Netherlands; [5]TATA Consultancy Services, Zaventem, Belgium; [6]Tilburg University, Department of Methodology and Statistics, Tilburg, Netherlands

Abstract

Adaptive learning systems (ALS) tailor the educational material to the level of the learners by continuously estimating their ability. By adaptively selecting practice items, ALS provide an optimal learning environment and decreases user dropout. While this adaptivity is favourable, it depends on the ability estimate of the student at a given time point, so if the estimates are inaccurate it is difficult to adapt item selection to the users' abilities. The initial uncertainty due to the limited information about user abilities is called the cold start and needs to be dealt with regardless of the method used for ability estimation. While the cold-start problem is the most pronounced when a user enters the system, it also manifests itself when there is a change in the learners' ability.

Due to the large-scale, adaptive and dynamic nature of ALS, they require innovative methods and algorithms to estimate ability on-the-fly. Recently a new algorithm called Urnings was developed which updates the estimates of ability and difficulty after every item response. We develop modifications of the Urnings algorithm to alleviate the cold-start problem by increasing the step size of the algorithm when a systematic change in the estimates is detected, and decreasing it when the estimates are relatively stable. We compare the modified algorithm with the algorithm with different fixed step sizes. Our results show that our modified algorithm moves away from the initial values faster, responds to sudden changes in ability better, and results in overall higher accuracy than the original algorithm.

Symposium title

Measuring changing abilities: Psychometrics for adaptive learning

Title

Introduction to computerized adaptive practice: Measurement challenges and solutions

Author(s)

Hanke Vermeiren[1,2], Abe D. Hofman[3,4], Maria Bolsinova[5], Wim Van den Noortgate[1,2], Han L.J. van der Maas[3]

[1]KU Leuven, Kortrijk, Belgium; [2]Imec research group Itec, Kortrijk, Belgium; [3]University of Amsterdam, Amsterdam, Netherlands; [4]Prowise, Amsterdam, Netherlands; [5]Tilburg University, Tilburg, Netherlands

Abstract

In this talk, we introduce computerized adaptive practice (CAP) and its implementation in adaptive learning environments for practicing math and language skills. CAP creates an optimal learning experience by providing practice items at the appropriate difficulty level. Adaptive item selection based on learner's ability has been suggested to positively impact motivation and learning outcomes. A prerequisite for adaptive item selection is precise assessments of ability levels. CAP is similar to computerized adaptive testing (CAT), which is aimed at selecting items with difficulty levels adapted to the ability of the individual thereby improving the efficiency of test-taking. However, CAP has important features which distinguish it from CAT. Several limitations of CAT hinder optimal application of the traditional IRT-based techniques to learning environments. For instance, CAT assumes known item difficulties, thereby requiring pretesting items. Learning environments often use vast item banks, making pretesting time-consuming. Furthermore, item selection techniques in CAT result in low levels of success making them less suitable in a learning environment due to high risk of drop-out. The pretesting problem is solved by using an algorithm based on the Elo rating system that allows for estimation of both item and learner parameters on the fly. Furthermore, CAP allows for easy (i.e., less informative, but more engaging) items by incorporating response times in the scoring of ability. We discuss the problems CAT techniques pose for adaptive learning environments as well as how the extended Elo algorithm implemented in CAP solves these issues.

Symposium title

Measuring changing abilities: Psychometrics for adaptive learning

Title

Modeling learning efficiency in adaptive learning environments with longitudinal IRT

Author(s)

Dries Debeer[1], Stefanie Vanbecelaere[2,3], Wim Van Den Noortgate[2,3], Bert Reynvoet[2], Fien Depaepe[2,3]

[1]Ghent University, Ghent, Belgium; [2]KU Leuven, Kortrijk, Belgium; [3]imec, Kortrijk, Belgium

Abstract

During the last decade, many governments and ed-tech companies have demonstrated an increased interest in digital personalised learning, which resulted in a variety of often game-like adaptive learning environments. However, there has been limited attention for the impact of these personalised learning technologies on children's learning efficiency. Does digital personalised learning, like popular claims insist, foster learning in young children? Such a question is not trivial from a methodological perspective, since the children's progress in the learning environment needs to be modeled appropriately. In this presentation we show how longitudinal IRT can be used to answer this question and present an empirical study which attempts to validate the beneficial impact of adaptive learning technology by analysing log-data from the Number Sense Game (NSG), an educational game that trains early numerical skills. In total, 81 children were randomly assigned to use either an adaptive or a non-adaptive version of the NSG in six sessions in a three-week period. Children's progress within and across sessions was modelled and compared between the two versions of the game. Regardless of the version of the NSG, children demonstrated progress within and across sessions. However, compared to the non-adaptive NSG, the progress across sessions was stronger in the adaptive NSG. These results provide empirical evidence that adaptive learning environments can improve learning efficiency in young children.

Symposium title

Measuring changing abilities: Psychometrics for adaptive learning

Title

Curious interactions between learners and Elo rating systems

Author(s)

Abe Hofman[1,2]

[1]University of Amsterdam, Amsterdam, Netherlands; [2]Prowise Learn, Amsterdam, Netherlands

Abstract

In this talk, I will share lessons learned from designing and maintaining a large adaptive learning platform (Prowise Learn). I will highlight several cases that show an interesting interaction between the Elo rating system and (unexpected) player behaviour. In these cases, self-reinforcing feedback loops between ratings and behaviour can have negative effects on the measurement properties and eventually break the system.
These results show the importance of monitoring our adaptive learning platform.

Symposium title

Measuring changing abilities: Psychometrics for adaptive learning

3.11 Parallel sessions 11h30–13h00 Auditorium 2

Symposium Overview

Modeling test-taking behavior: Moving from pure nuisance to relevant substantive phenomena

Author(s)

Nico Remmert[1], Ulf Kröhne[2], Marek Muszyński[3], Susana Sanz Velasco[4], Steffi Pohl[1]

[1]Freie Universität Berlin, Berlin, Germany; [2]DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation, Frankfurt am Main, Germany; [3]Polish Academy of Sciences, Warsaw, Poland; [4]Universidad Autónoma de Madrid, Madrid, Spain

Abstract

Traditionally, in the assessment of psychological constructs, missing values, differences in speed, and cursor movements have been considered as nuisance that need to be controlled for. Thus, modelling these phenomena predominantly aimed at controlling for confounding sources, reducing bias or enhancing estimation precision. However, these phenomena are not only nuisances, but may also represent substantively relevant phenomena that are informative of person characteristics. In this symposium we will present research that demonstrates the utility of these phenomena as a substantial source to assess underlying processes, behavior, and person characteristics. Using different indicators of test-taking behavior in different substantive areas, the symposium shows that a lot can be gained by moving from considering behavior when approaching tests or questionnaires as pure nuisances to relevant substantive phenomena. In the contributions of this symposium, a) disengagement during test taking in large-scale assessments is identified using action sequence data when solving a task, b) it is shown that a set of log-data indices validate experimentally induced satisficing in online surveys, c) that missing values provide relevant information about persons abilities that are predictive for later educational outcomes, and d) that missing values and response time data can be used in an innovative way to move from self-reports to measuring actual avoidance behavior.

Title

Predictive validity of missing values for later educational outcomes

Author(s)

Susana Sanz[1], Esther Ulitzsch[2], Carmen García[1], Ricardo Olmos[1], Steffi Pohl[3]

[1]Universidad Autónoma de Madrid, Madrid, Spain; [2]2IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Kiel, Germany; [3]Freie Universität Berlin, Berlin, Germany

Abstract

Performance in a test does not only depend on cognitive skills, but also on test-taking strategies. While missing values are generally considered as nuisances, we investigate in which way these are informative of examinee's success in life. In low-stakes contexts, information on how examinees approach test items could help understand how an examinee manages tasks in general, and ultimately, it could be related to later success. In this sense, extracting information from missing values in tests, considering item omissions, and not-reached items could be a powerful resource. Using competence test data from various competence domains in the longitudinal German National Educational Panel Study, we investigated a) which person variables may explain the occurrence of missing values and b) in which way missing values are informative (over and above competence) for educational outcomes measured up to 6 years after the competence assessment. We incorporated the missing tendency due to omission (Holman & Glas, 2005) and not reaching the end of the test (Rose et al, 2010) in the competence model and estimated both, a) correlations of the missing propensities of explaining variables and b) the predictive validity of missing propensities to later educational outcomes. We found that self-concept and educational aspirations partly explain the presence of missing values. Most interestingly we found that missing propensity predicted final grades, employment income, and grade repetition. These results support the idea that how competence test items are approached can provide valuable additional information on general abilities of examinees that are relevant for real life.

Symposium title

Modeling test-taking behavior: Moving from pure nuisance to relevant substantive phenomena

Title

A void as a reflection of avoidance: Using missing responses and response times to assess avoidance behavior

Author(s)

Nico Remmert[1], Silia Vitoratou[2], Jane Gregory[3], Svetlana Shinkareva[4], Sewon Oh[4], Robert Krause[5], Steffi Pohl[1]

[1]Freie Universität Berlin, Berlin, Germany; [2]King's College London, London, United Kingdom; [3]University of Oxford, Oxford, United Kingdom; [4]University of South Carolina, Columbia, USA; [5]University of Kentucky, Lexington, USA

Abstract

Avoidance is a central behavior related to psychopathology. It adversely contributes to the maintenance of clinical symptoms and has an impact on treatment outcomes. Current assessments of avoidance behavior mainly use self-reports. As compared to the measurement of actual behavior, self-reports are usually less valid. We propose an approach for the measurement of actual avoidance behavior by making use of psychometric models for response times and missing values. We assess actual avoidance of adverse auditory stimuli in a computerized test as natural reflections of behavioral avoidance. In our approach, we model the occurrence of avoidance by relying on psychometric models for missing values; the related time until avoidance occurs is modelled by psychometric models for response time data. We demonstrate our modeling framework with a preliminary example of sound avoidance in a sample of individuals suffering from misophonia (i.e., selective sound intolerance). The results and implications are discussed against the background of psychopathology and psychometrics. We argue that our assessment as well as the modeling framework provides a promising and efficient way to move from self-reports to the measurement of actual behavior in psychopathology.

Symposium title

Modeling test-taking behavior: Moving from pure nuisance to relevant substantive phenomena

Title

On the Structure of Disengagement in Questionnaires and Cognitive Tests with an Example from PISA

Author(s)

Ulf Kroehne[1], Frank Goldhammer[1,2]

[1]DIPF | Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany; [2]Centre for International Student Assessment (ZIB), Frankfurt am Main, Germany

Abstract

Disengaged responses are a well-known threat to the validity of score interpretation in cognitive assessments and questionnaires. Various indices have been developed to detect disengaged respondents (e.g., self-report measures) or responses (e.g., based on response consistency, missing value pattern, or response times). This paper investigates the relationship between process indicators for disengaged responding across different assessment parts (i.e., tests and questionnaires). First, we theoretically compare the concepts of response-time-based identification of rapid guessing (in tests) and its counterpart of rapid responding (in questionnaires), describe the assumption regarding the response processes, and summarize the available validity evidence for these process indicators as disengagement measures. Using empirical data from the PISA 2015 assessment of selected countries, we then investigate the structure and dimensionality of the two process indicators. For validating the process indicators, we investigate their relationship to student background variables (gender, socio-economic status), cognitive competencies, consistency measures for questionnaires (i.e., measures for careless insufficient effort responding), and missing value patterns (e.g., number of omitted and not-reached items). Based on the empirical example, we finally investigate the substantive question of whether rapid responding in the questionnaire results from low reading competence, taking rapid guessing in the cognitive tests into account. In the closing discussion, recommendations are formulated on how future psychometric assessments can incorporate additional log data to improve the interpretation of engagement-related process indicators.

Symposium title

Modeling test-taking behavior: Moving from pure nuisance to relevant substantive phenomena

3.12 Parallel sessions 11h30–13h00 Auditorium 3

Title

Thinking beyond the mean - how to escape a paradigm of averages

Author(s)

Daniel Gotthardt[1], Christoph Naefgen[2], Anne Reinarz[3]

[1]Department of Social Sciences, Universität Hamburg, Hamburg, Germany; [2]Faculty of Psychology, FernUniversität in Hagen, Hagen, Germany; [3]Department of Computer Science, Durham University, Durham, United Kingdom

Abstract

Quantitative behavioral and social scientists increasingly try to grasp beyond the reach of the dominant paradigm of averages, frequently by focusing on progressively complex and heterogeneous conditional mean estimation. Meanwhile, conditional dispersion is ubiquitous in experimental, survey and natural data and variability is implicitly specified in a multitude of social, political, economic and psychological theories. However, variability is usually only seriously considered in qualitative research, and a few select subfields like inequality research and group emotions. Due to a strong focus on conditional expectations in the statistical training of researchers, we often don't even considerthinking beyond the mean.
We propose utilizing variability more in all stages of research: Deliberately theorizing about dispersion, using distributional statistical models and interpreting variability as meaningful information instead of mere noise. In support of methodological efforts to promote theory formalization and reflection on estimands, we present an approach to expressing theoretical assumptions about variability making them empirically tractable.We use a multilevel perspective to clarify the meaning of variability in context and develop a preliminary overview of categories of mechanisms implying conditional dispersion. We use Monte-Carlo simulations to show how quantile and variance based methods perform in some semi-realistic scenarios. Lastly, we showcase the utility of variability approaches using studies from cognitive psychology and sociology of religion, two fields where quantitative studies of variability are traditionally scarce.
Our goal is embedding more technical discussions regarding distributional modeling in a metatheoretical framework in order to provide empirical researchers with easier access to methodological insights.

Oral presentations session title:

Multilevel Analysis

Title

How Should Pretest Measures be Included in Multilevel Models When Examining the Effects of Teacher- or School Variables on Learning?

Author(s)

Carmen Köhler[1], Marie-Ann Sengewald[2], Steffen Zitzmann[3], Peter Edelsbrunner[4]

[1]DIPF, Frankfurt, Germany; [2]Lifbi, Bamberg, Germany; [3]Uni Tübingen, Tübingen, Germany; [4]ETH Zürich, Zürich, Switzerland

Abstract

Researchers commonly use a multilevel modelling approach to examine the effects of teacher or school variables on students' learning outcomes. To estimate such effects, educational assessment studies oftentimes contain a pretest measurement of the outcome variable that can be included as a covariate in the multilevel model. The covariate can be incorporated either solely at the student level (L1) or at both the student and class (or school) level (L2). In practice, this is handled differently across applied studies. From a methodological viewpoint, various potential factors that might bias the estimated effect of teaching quality need to be considered in order to answer the question of whether the pretest should additionally be included at L2: omitted confounders, contextual effects, and L2-endogeneity. In a simulation study, we examine for which of these factors—and especially in their interplay—pretest measures should be included at L2. Our findings show that, without confounding between the L2 pretest and the teaching quality variable, without contextual effects and without L2-endogeneity, both models lead to an unbiased estimate of the effect of teaching. When confounding is present and a contextual effect exists (but no L2-endogeneity), the L2 pretest should be included in the model. When confounding and L2-endogeneity are present (but no contextual effect), the L2 pretest should not be included in the model. When all three exist simultaneously, both models show bias. We discuss which of the three factors are likely to occur in various applied scenarios, and give recommendations for applied researchers.

Oral presentations session title:

Multilevel Analysis

Title

Consequences of Hierarchical Data Structures for the Estimation of Plausible Values: An Application of Multilevel Modeling in Educational Large-Scale Assessments

Author(s)

Eva Zink[1], Sabine Zinn[2], Timo Gnambs[1]

[1]Leibniz Institute for Educational Trajectories, Bamberg, Germany; [2]German Institute for Economic Research, Berlin, Germany

Abstract

Educational large-scale assessments (LSAs) measure domain-specific competences and general cognitive abilities to monitor educational processes within and between countries. In these studies, competence scores are typically represented by a set of plausible values (PVs) that allow the analysis of latent effects. So far, an unresolved challenge in the estimation of PVs is the multilevel structure that arises when students (level 1) are clustered in different school contexts (level 2). Current practices involve ignoring the hierarchical data structure or including cluster-specific mean scores in the background model for the PV estimation. However, both approaches might bias substantial analyses if the generated PVs do not appropriately reflect the multilevel structure. Therefore, a Monte Carlo simulation evaluated the consequence of either ignoring or including the hierarchical data structure in the background model of the PVs estimation. As an alternative to current practice, we compare whether including a random effect in the background model to account for the hierarchical data structure improves parameter accuracy. The simulation study mimicked realistic data in LSAs and evaluated the performance of the different approaches in an experimental 3 x 3 design formed by the PV estimation model (ignoring, cluster scores, random effects) and analysis model (ignoring, cluster, effects, random effects). We simulated responses of 4000 respondents nested in 20 clusters to 20 items conforming to the one-parametric item response model. The bias, relative bias, and root mean squared error of parameters resulting from regressing the PVs on a continuous and a binary predictor were our main evaluation criteria.

Oral presentations session title:

Multilevel Analysis

Title

Removing common-source variance from aggregated multilevel data: Possible and worth the hassle?

Author(s)

Jonas Lang

University of Exeter, Exeter, United Kingdom

Abstract

Applied researchers are frequently concerned about the possible influence of common method biases on inferences (Campbell & Fiske, 1959; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). This concern is particularly frequently voiced for self-report data from organizations like employee engagement surveys. This type of data is frequently collected to gain insights about organizational units like teams, departments, or branches of organizations and their leaders. One characteristic of aggregated multilevel data that is frequently overlooked is its lower-level multi-source nature with multiple raters rating the same target. In this presentation, I study the possibility to remove common-rater variance in aggregated multilevel constructs through the use of a multidimensional cross-classified mixed-effect model with items for multiple constructs nested in raters and units called the unit-rater multilevel model (URM). Analyses using the lme4 package (Bates et al., 2022) in R revealed that differences between the corrected and the uncorrected/raw coefficients in a classic organizational dataset from the literature. A simulation study further showed that common-source bias—if uncorrected—indeed biases ICC1 values downward and inter-construct correlations either upward or downward depending on the circumstances. The simulation study also showed that the URM was capable of significantly reducing this bias at small sample sizes and effectively removed it at larger sample sizes. An important implication of the study for future organizational research is that multilevel survey data should be viewed as a type of multi-source data.

Oral presentations session title:

Multilevel Analysis

3.13 Parallel sessions 11h30–13h00 Auditorium 4

Title

Prospectively monitoring the intensity and variability of emotions using statistical process control

Author(s)

Marieke Schreuder[1], Evelien Schat[1], Arnout Smit[2], Evelien Snippe[3], Eva Ceulemans[1]

[1]KU Leuven, Leuven, Belgium; [2]Vrije Universiteit, Amsterdam, Netherlands; [3]UMCG, Groningen, Netherlands

Abstract

Statistical process control (SPC) can track changing mean levels in repeatedly assessed emotions (e.g., feeling restless) in real-time, which has proven useful for the early detection of depression. This preregistered study investigated whether depression detection improves when SPC is used to not only monitor mean levels, but also changing standard deviations (SDs) of emotions. We investigated 41 formerly depressed adults who monitored their emotions five times a day for four consecutive months. During the study, 22 participants experienced depression recurrence. We used SPC to detect whether these recurrences were anticipated by warning signs (i.e., changing means and SDs) of five emotions (positive and negative affect with high or low arousal; repetitive negative thinking). For each participant and each emotion separately, we derived control limits based on assessments in the first four weeks and subsequently monitored whether and when exponentially weighted moving means/SDs crossed these control limits. Matthews correlation coefficient (MCC), with p-values based on permutation tests, was used to assess predictive performance. We found that (1) SD-based warning signs in most emotions predict recurrent depression with high specificity (p-values < 0.002), (2) mean- and SD-based warning signs show little overlap (p-values > 0.05), and (3) the detection of depression generally improved when monitoring both mean- and SD-based warning signs (p-values < 0.015, except for PA low arousal: $p=0.085$). Thus, real-time detection of changing SDs by means of SPC may advance the early detection of depression.

Oral presentations session title:

Intensive Longitudinal Data 2

Title

Extracting personalized latent dynamics over time using the multilevel hidden Markov model in social and behavioral processes: the R CRAN package mHMMbayes and empirically based guidelines on sample size requirements

Author(s)

Emmeke Aarts, Sebastian Mildiner Moraga

Utrecht University, Utrecht, Netherlands

Abstract

The multilevel (also known as mixed or random effects) hidden Markov model - a generalization of the hidden Markov model (HMM) - is a promising vehicle to investigate latent dynamics over time in social and behavioral processes in intense longitudinal data. The multilevel HMM is tailored to accommodate data of multiple individuals simultaneously, allowing for heterogeneity in the model parameters (transition probability matrix and conditional distribution), while estimating one overall HMM. Hence, the multilevel framework facilitates the study of individual-specific trajectories and the study of individual differences.
An open-source implementation of the multilevel hidden Markov model is provided by the R CRAN package mHMMbayes. The model can be fitted on multivariate data with a categorical or normal (i.e., Gaussian)distribution, and include individual level covariates (allowing for e.g., group comparisons on model parameters). Parameters are estimated using Bayesian estimation utilizing the forward-backward recursion within a hybrid Metropolis within Gibbs sampler. The usefulness of the multilevel HMM and mHMMbayes is illustrated using an example on ESM data on five cognitive, affective, and behavioral (CAB) factors, in which the multilevel HMM unveiled four distinctive CAB crisis states and patient individual trajectories herein.
In addition, we provide guidelines on sample size requirements – currently still lacking for typical social and behavioral data in combination with the multilevel HMM. The guidelines are based on extensive simulation studies and are driven by the complexity of the data and the study objectives of the practitioners.

Oral presentations session title:

Intensive Longitudinal Data 2

Title

Enabling analytical power calculations for multilevel models with autocorrelated errors through deriving and approximating the information matrix

Author(s)

Ginette Lafit, Richard Artner, Eva Ceulemans

KU Leuven, Leuven, Belgium

Abstract

To unravel how within-person psychological processes fluctuate in daily life, and how these processes differ between persons, intensive longitudinal (IL) designs in which multiple participants are repeatedly measured, have become popular. Commonly used statistical models for those designs are multilevel models with autocorrelated errors. To then examine substantive hypotheses of interest, statistical hypothesis tests are conducted for the effects of interest in the fitted multilevel model. An important question in the design of such IL studies concerns the determination of the number of participants and the number of measurements per person needed to achieve sufficient statistical power. Recent advances in computational methods and software enable the computation of statistical power using Monte Carlo simulation. Unfortunately, this approach is highly computationally intensive. We, therefore, derive analytical formulas for statistical power in multilevel models with AR(1) within-person errors. Analytic expressions are obtained via asymptotic approximations for the unknown quantities in the information matrices of the fixed effects. To validate this analytical approach, we perform a series of simulations to compare its performance to the simulation-based approach. The approaches perform similarly thereby making the analytic approach a viable option for researchers that can significantly reduce the computational burden.

Oral presentations session title:

Intensive Longitudinal Data 2

Title

Intensive Longitudinal Measurement: How precise can we be and when?

Author(s)

Joran Jongerling

Tilburg University, Tilburg, Netherlands

Abstract

Science is shifting towards studying processes as they unfold over time, for example using experience sampling methodology (ESM), as opposed to investigating one (or a handful) of snapshots from these processes. Unfortunately, approaches to account for measurement error (e.g., with factor models) are not yet routinely applied to the intensive longitudinal data (ILD) obtained using ESM and related methods. Not because researchers are unaware that failing to account for measurement error in their analyses threatens the validity and practical usefulness of findings but because properly modeling measurement error in ILD comes with unique challenges. For example, due to the high frequency of measurements (e.g., six times per day), ESM questionnaires must not be too long. This poses a problem for factor models—the "gold standard" for modeling measurement error in cross-sectional data—which need several items for each construct of interest. Without enough items, factor models might actually bias results more than simpler models that ignore measurement altogether. In this talk, I will take an extensive look at different methods to include measurement error in ILD to determine which approaches are feasible under realistic conditions, and to highlight how they are effected by different design choices. I will focus on dynamic factor modeling, a method introduced by Schuurman & Hamaker (2019) that adds a measurement-error term to composite scores but does not include an actual measurement model, and models that analyze uncorrected composites. I will provide guidelines on which methods work best under which circumstances.

Oral presentations session title:

Intensive Longitudinal Data 2

3.14 Parallel sessions 11h30–13h00 Lecture room 1.2

Title

Using Multiple Group SEM in place of Mediation Models for Experimental Research

Author(s)

Jonathan Helm

San Diego State University, San Diego, USA

Abstract

Mediation models (MMs) are a subset of structural equation models (SEMs) which estimate the mediating effect of a variable (M) between a predictor (X) and outcome (Y). For experimental research, MMs estimate the degree to which the treatment (e.g., experimental vs. control condition) affects the outcome via a mediator. However, MMs assume the following are equal across treatment conditions: (1) the variance of the mediator, (2) the residual variance of the outcome, and (3) the effect of the mediator on the outcome. Furthermore, if latent variables are used to measure (M) and/or (Y), then measurement invariance of the latent variables is typically assumed across the treatment conditions. The results from MMs (e.g., estimate of the indirect effect) will be inaccurate (e.g., heightened bias and Type-1 error) to the extent these these assumptions do not hold. This talk will show how to test and relax these assumptions using a multiple group structural equation model. More specifically, it will be shown that the MM may be written as a constrained multiple group SEM (with constraints equal to assumptions of the MM). A new estimate for the indirect effect will be derived based on the multiple group SEM, which exactly equals the indirect effect from the MM when all assumptions hold. The multiple group SEM can also test for measurement invariance for latent measures ofM and Y.The results are demonstrated with a small simulation, and emphasized using an empirical example.

Oral presentations session title:

Multigroup and Meta-Analytic SEM

Title

Mixture multigroup SEM for comparing structural relations among many groups

Author(s)

Andres Felipe Perez Alonso[1,2], Jeroen Vermunt[1], Yves Rosseel[3], Kim De Roover[2]

[1]Tilburg University, Tilburg, Netherlands; [2]KU Leuven, Leuven, Belgium; [3]Ghent University, Ghent, Belgium

Abstract

Social scientists often examine the relationships between two or more latent variables or constructs, and Structural Equation Modeling (SEM) is the state-of-the-art for doing so. When comparing these structural relations among many groups, they likely differ across the groups. However, it is equally likely that some groups share the same relations, and that clusters of groups emerge in terms of the relations between the latent variables. For validly comparing the latent variables' relations among groups, the measurement of the latent variables should be invariant across the groups (i.e., measurement invariance), whereas often at least some measurement parameters differ across the many groups. Restricting these measurement parameters to be equal across groups, when they are not, causes the structural relations to be estimated incorrectly and thus invalidates their comparison. Therefore, to capture differences and similarities in structural relations while accounting for the reality of measurement non-invariance, we propose mixture multigroup SEM (MMG-SEM). MMG-SEM obtains a clustering of groups focused entirely on the structural relations by making them cluster-specific, while allowing for the measurement parameters to be (partially) group-specific. In this way, MMG-SEM disentangles differences in structural relations from differences in measurement parameters. We present an expectation-maximization estimation procedure, built around the R-'lavaan', as well as an evaluation of MMG-SEM's performance in terms of recovering the group-clustering and the group- and cluster-specific parameters.

Oral presentations session title:

Multigroup and Meta-Analytic SEM

Title

The Impact of Non-Normality on Causal Effects in Path and Multigroup Structural Equation Models with Ordered Categorical Variables

Author(s)

Benedikt Lugauer, Jana Holtmann

Leipzig University, Leipzig, Germany

Abstract

In the social and behavioral sciences, assessing the causal effect of a treatment on an outcome is a central research goal. Average, conditional, direct and indirect effects are common measures to estimate these causal effects using regression or multigroup structural equation models (SEM). Ordered categorical response formats are common for survey items in these fields, which are either modeled as continuous (using maximum likelihood; ML), ignoring their actual distribution, or treated as the result of a discretization process (using diagonally weighted least squares estimation; DWLS). While ML relies on the assumed multivariate normality of the observed variables, DWLS assumes that ordered categorical variables result from discretizing an underlying normal variable. Previous studies investigating the effect of violations to the normality assumption relied on the Vale-Maurelli approach, which has been shown to produce data equivalent to simulating from a multivariate normal random vector (Grønneberg & Foldnes, 2019). This simulation study examines the effect of non-normality on estimating treatment effects in path and multigroup SEM with ordered categorical variables based on the newly developed vine copula approach, Vine-to-Anything (VITA; Grønneberg, Foldnes, and Marcoulides, 2022). We investigate the effect of complex multivariate distributions by varying the marginals and the bivariate copulas independently from each other. Results suggest that the amount of bias in the effect estimates critically depends on the combinations of multivariate distributions (copulas) and marginal distributions in the treatment and control groups in multigroup SEMs.

Oral presentations session title:

Multigroup and Meta-Analytic SEM

Title

In Between Methods: Evaluating Approaches for Individual Participant Data Meta-Analytical Structural Equation Modeling

Author(s)

Lennert Groot, Kees Jan Kan, Suzanne Jak

University of Amsterdam, Amsterdam, Netherlands

Abstract

Researchers conducting meta-analytical structural equation modeling (MASEM) using raw data have several analysis options to choose from. Cluster-robust estimation, two-level structural equation modeling (SEM), multi-group SEM, meta-analysis of path coefficients, and One-Stage MASEM (OSMASEM) are some of these options. Two-level SEM explicitly separates effects at the within-study level from the between-study level, multi-group SEM and OSMASEM look at the within- study effects, while cluster-robust method estimates an overall path coefficient, which essentially is a mix of within-study and between-study effects. Of these, cluster-robust estimation is often used in practice. A comparison of these methods using real-world data, however, shows that cluster-robust estimates deviate from results of other methods. Simulations using a factor model have shown that cluster-robust estimation may not always be free of bias. This study evaluates bias in parameter estimates and standard errors of MASEM methods with raw data in the context of path analysis, using simulated data. We varied equality of path models and coefficients over the within-study and between-study level, number of primary studies being meta-analyzed, and amounts of study-level heterogeneity. Results are expected to show that cluster-robust estimation method yields significant bias in estimated path coefficients in certain conditions.

Oral presentations session title:

Multigroup and Meta-Analytic SEM

3.15 Parallel sessions 11h30–13h00 Lecture room 1.3

Title

Statistical inference for agreement between multiple observers on a binary scale

Author(s)

Sophie Vanbelle

Maastricht University, Maastricht, Netherlands

Abstract

Agreement studies with binary outcomes are frequent in psychological and medical sciences. For example, in clinical decision making, disagreement between psychiatrists can lead to different treatments for the patient. In the presence of two observers or two repeated measurements (R>2), simple agreement measures such as the proportion of agreement and specific agreement measures (e.g., positive agreement) are popular. These measures were generalised to account for more than two observers or repeated measurements (R>2) by de Vet and al. (2017) but statistical inference was only proposed on an empirical basis. In this talk, we propose statistical inference in the frequentist and the Bayesian framework. In the frequentist framework, an analytical formula for the large sample variance of the agreement measures is derived using the delta method and used to construct a Wald confidence interval. In the Bayesian framework, statistical inference is based on the results of Bloch and Watson (1967). The posterior distribution of the agreement measures can be approximated analytically and does not require any specific Bayesian statistical software. Based on the study of the statistical properties of the confidence and credibility intervals, guidelines to make statistical inference are provided. Further, in the frequentist framework, analytical formulas are derived to determine the number of participants needed in an agreement study for a given number of observers. Finally, a Shiny application and a R package are presented to facilitate the use of the proposed statistical techniques.

Oral presentations session title:

Rater Agreement and Reliability

Title

How to determine sample size for the Intraclass Correlation Coefficient in the one-way ANOVA model

Author(s)

Dipro Mondal[1], Sophie Vanbelle[1], Math Candel[1], Alberto Cassese[2]

[1]Maastricht University, Maastricht, Netherlands; [2]The University of Florence, Florence, Italy

Abstract

Reliability evaluation is a crucial part in any quantitative study. The reliability of measurement instruments providing continuous outcomes is usually assessed by the Intraclass Correlation Coefficient (ICC). In planning a reliability study where some participants are measured repeatedly by a single rater or device, the number of participants, n, and the number of measurements per participant, k, need to be determined. When the order of the measurements is exchangeable, the data can be modelled by a one-way ANOVA model. This presentation gives an overview of the different approaches to determine n for a fixed k, given a specific value of the ICC under the one-way ANOVA model. Sample size determination approaches under this model are based on the limits of a confidence interval for the ICC. Although eight different confidence interval methods can be identified in the literature, Wald confidence interval with large sample variance approximation developed by Fisher remains the most commonly used despite its well-known poor statistical properties. Therefore, a first objective of this work is to compare the statistical properties of all identified confidence interval methods - including the ones overlooked in previous studies. A second objective is to develop a general procedure for determining n, since a closed form formula for sample size determination is not always available. This new procedure is implemented in an in-house Shiny/R app. Finally, we provide guidelines for choosing an appropriate sample size determination method when planning a reliability study.

Oral presentations session title:

Rater Agreement and Reliability

Title

Reliability for multilevel data: A correlation approach

Author(s)

Tzu-Yao Lin[1,2], Francis Tuerlinckx[2], Sophie Vanbelle[1]

[1]Maastricht University, Maastricht, Netherlands; [2]KU Leuven, Leuven, Belgium

Abstract

Studying the reliability of a measurement instrument is essential. Despite the awareness of the problem of measurement errors in psychology and medicine and the various reliability coefficients that have been proposed, research on reliability for multilevel data, which ubiquitously exist in observational studies, remains limited. Two recent papers (Schönbrodt et al., 2022; ten Hove et al., 2022) address how to quantify reliability in multilevel settings based on generalizability theory. Specifically, ten Hove et al. (2022) defined between-cluster and within-cluster interrater intraclass correlation coefficients for multilevel designs where subjects or raters are nested within clusters. Schönbrodt et al. (2022) also defines reliability coefficients at between-cluster and within-cluster (i.e., between-subject) levels for designs where dyad members nested in couples are assessed numerous times daily over a number of days. Nevertheless, both approaches give inconsistent results regarding their definition of cluster-level reliability. In this paper, we propose an alternative approach to define reliability coefficients that are based on calculating the expected correlation between repeated measurements (Molenberghs et al., 2007; Vangeneugden et al., 2005). We will compare our approach with that of Schönbrodt et al. (2022) and ten Hove et al. (2022) and explain the differences between the three approaches in two common nested data structures: (1) raters crossed with both subjects and clusters, but subjects are nested within clusters and (2) raters nested within both subjects and clusters. To conclude, we will provide guidelines to measure reliability in multilevel data structures.

Oral presentations session title:

Rater Agreement and Reliability

Title

Comparing Interrater Reliability Estimates across Estimators and Different Incomplete Observational Designs

Author(s)

Debby ten Hove

Vrije Universiteit Amsterdam, Amsterdam, Netherlands

Abstract

Interrater reliability (IRR) is imperative for observational research. It involves the degree to which observations are independent of raters, and serves as an indicator for measurement precision and (loss of) statistical power in subsequent analyses. To distribute the workload across raters, many observational studies use an incomplete observational design in which the observers vary across subjects. Some studies use randomly sampled raters for each subject, others use randomly sampled pairs of raters for each subject, and yet others vary one or more of the raters across subjects while keeping one rater constant for all subjects. Traditional estimation methods for ICCs, which are used to investigate the interrater reliability (IRR) of observations, cannot handle such incomplete designs. I therefore conducted a simulation study to compare several novel estimation methods of ICCs for IRR under various data-generating conditions: ICCs based on the variance decomposition of (1) Markov chain Monte Carlo (MCMC) estimation of hierarchical linear models, (2) maximum likelihood estimation (MLE) estimation of a latent variable model, and (3) MLE of a random-effects models. I will pay special attention to the effect of the different observational designs on the point- and interval estimates of the ICCs, and present software for the MLE-approach, which yields the most precise point- and interval estimates across conditions.

Oral presentations session title:

Rater Agreement and Reliability

3.16 Poster session 3 14h00–15h00

Title

Measurement Invariance of the Problematic Use of Social Networks Questionnaire between different User Profiles

Author(s)

Covadonga González-Nuevo, Jaime García-Fernández, Álvaro Postigo, Marcelino Cuesta

University of Oviedo, Oviedo, Spain

Abstract

Background: There are numerous instruments for assessing problematic Social Network Sites (SNS) use. These questionnaires assume that the construct of problematic SNS use is measured across different genders and social network users. However, as far as we know, no study tested whether the questionnaire used met the measurement invariance between different users of each social network and gender. In this context, the aim of this study was to test the measurement invariance of the Problematic Use of Social Networks Questionnaire (UPS) for Facebook, Instagram and TikTok users and between genders. Method: 1,887 participants (69.3% female) with an age range between 18 and 78 years (M = 25.78; SD = 12.31) answered an online questionnaire conditional on being active users of Facebook, Instagram or TikTok. Different levels of invariance (configural, metric and scalar) were tested for TikTok users, Instagram users, Facebook users and for each gender. Results: The UPS showed measurement invariance across user profile and gender. Conclusions: This study allows us for the first time to test the measurement invariance of a questionnaire (i.e. UPS) for assessing problematic social media use across different user types.

Title

Psychometric evaluation of the German Big Five Inventory-2 using a multi-method, multi-context approach

Author(s)

Dora Leander Tinhof, Axel Mayer

Universität Bielefeld, Bielefeld, Germany

Abstract

The Big Five personality traits are widely used and structure personality into five distinct dimensions, which have been consistently found across a wide range of contexts. Accurate and reliable measurement of psychological constructs across different contexts plays a crucial role in the replicability of scientific research. The digital world represents a context which has gained increasing importance during the COVID-19 pandemic. Given its significant differences from the offline world, it is essential to also investigate the applicability of the Big Five's dimensional structure in an online context.

Therefore, the current study investigates the psychometric properties of the German adaption (Danner et al., 2019) of the Big Five Inventory-2 (BFI-2; Soto & John, 2017) using data from the first measurement wave of a larger, longitudinal study. The BFI-2 introduces three facets for each of the five trait domains and provides both self- and other-report questionnaires. Using confirmatory factor analysis, measurement invariance is evaluated across all four combinations of methods (self- & other-report) and contexts (offline & online). First, configural invariance is assessed separately within each of the five trait domains as well as for the overall five-dimensional factor structure. Subsequently, measurement invariance tests up to residual (strict) invariance are conducted. Final model structures and implications are discussed. Cronbach's Alpha and McDonald's Omega are reported on facet and domain level, as well as relevant item reliabilities and loadings. Additionally, self-other agreement, substantial correlations with variables of interest and initial findings of longitudinal analyses, including retest-reliabilities and trait differences across contexts, are presented.

Title

Bayesian Generalized Method of Moments approach for estimating Rank Preserving Models: A flexible approach for causal mediation analysis.

Author(s)

Roberto Faleh, Holger Brandt

University of Tübingen, Tübingen, Germany

Abstract

Mediation analysis is a fundamental tool in empirical sciences, particularly in medical and social sciences, where intermediate variables play a crucial role in understanding treatment efficacy. The classical regression-based mediation analysis method proposed by Baron and Kenny has been criticized for its restrictive assumption of no-unmeasured-confounder, which requires that all confounders have been measured and incorporated into the model. This assumption can be difficult to satisfy in practice and violations lead to spurious results and make causal interpretation challenging.

One of the most prominent models relaxing the assumption of no-unmeasured-confounder is the Rank Preserving Model (RPM), introduced by Ten Have and colleagues. The RPM assumes that unobserved confounders do not interact with treatment or mediators. This assumption is often more plausible than the no-unmeasured-confounders assumption making the model relevant in less confining theoretical and empirical circumstances.

To further generalize the model and weaken the assumptions required by the RPM, Zheng and colleagues proposed a more flexible model that can handle multiple mediators and multilevel interventions.

However, models using the RPM assumption have not been used extensively due to low power and inefficiency in many scenarios. The Bayesian Generalized Method of Moments (GMM) is proposed as a solution to improve the power and flexibility of the RPM.

The Bayesian Generalized Method of Moments has several advantages over classic frequentist approaches, including the possibility to directly derive standard errors for estimates as well as higher power, robust, and unbiased estimation of the model's parameters.

Title

Enhancing Predictive Cut Scores in Higher Education Enrollment with Explainable Machine Learning Algorithms

Author(s)

Tuo Liu, Andreas Frey

Goethe University Frankfurt, Frankfurt am Main, Germany

Abstract

The competitive landscape of higher education enrolment requires data-driven decision-making by enrolment managers supported by machine learning techniques. However, previous machine learning-based systems have typically only provided a predictive cut score, lacked expandability and failed to communicate the level of confidence in the predictions. As a result, erroneous decision-making may ensue. To address this limitation, we propose a novel machine-learning framework for determining the cut score for the threshold of offering admission, which incorporates four explicable machine-learning algorithms (Generalized Linear Models, Elastic Net, Decision Trees and Random Forests). The proposed framework uses historical data to determine the probability that an applicant will enrol in the program once accepted. Based on the annual enrolment target, an optimal cut score is determined using the mathematical expectation of the probability of enrolment. Unlike previous machine learning approaches, the proposed framework provides information on the explainable variable importance and confidence level of the cut score, supporting rational decision-making. The enhanced machine learning-based cut score method can potentially assist college admissions officers in efficiently using educational resources.

Title

Benefits of multinomial processing tree models with discrete and continuous variables.

Author(s)

Anahí Gutkin, Manuel Suero, Juan Botella

Social Psychology and Methodology, Universidad Autónoma de Madrid, Madrid, Spain

Abstract

Cognitive architecture can be translated mathematically through models, among them multinomial process tree models, MPTs, that allow estimating the probabilities of latent processes given observed responses categories. Recently, MPTs procedures were proposed to jointly model discrete and continuous variables (MPT-DC) (Heck & Erdfelder, 2016, 2018; Klauer & Kellen, 2018). In this study, we analyze the benefits of MPT-DC. We employed the signal detection theory, SDT, and the two-high threshold, 2HT, models in a recognition memory paradigm where confidence levels and response times are measured. Thus, MPT-DCs have been fitted including these two variables. Based on experimental and simulated data, we have found that jointly modeling relevant categorical and quantitative variables: (a) reduces the standard error of estimates; (b) correct model selections are increased; and (c) detect interactions that classical versions of the models don't. In conclusion, the MPT-DCs models provide methodological and substantive advantages in the disentanglement of cognitive processes.

Title

Measuring Pro-Environmental Behaviour: convergent validity, internal consistency, and respondent experience of existing instruments

Author(s)

Berre Deltomme[1], Karen Gorissen[2], Bert Weijters[1]

[1]Ghent University, Ghent, Belgium; [2]Vrije Universiteit Amsterdam, Amsterdam, Netherlands

Abstract

The rising attention to the influence of human behaviour on climate change and environmental decline has led to an increase in studies that measure Pro-Environmental Behaviour (PEB) as a predictor, a covariate, or an outcome variable. To this end, (validated) self-report scales have traditionally been the main measurement tool, but lately, several experimental instruments have been developed to measure PEB. Measurement instruments that are considered to measure the same construct should provide consistent results, i.e., they should show high convergent validity. However, it is not clear whether substitute measures for PEB show this necessary convergent validity and how they relate to each other in terms of internal consistency and respondent experience. To gain insight herein we investigated thirteen validated self-report scales and three experimental tasks on their psychometric qualities (i.e., validity and internal consistency) and respondent experience. The results show that, in general, convergent validity is lacking, showing that the measurement instruments cannot be considered equivalent. As for respondent experience, the experimental tasks are the most time-consuming, are perceived as most fatiguing, and are most sensitive to multitasking. The self-report scales are most sensitive to social desirable responding and acquiescence bias. We provide ideas on how convergent validity might be heightened. Next to that, we discuss a guideline that might help researchers in selecting the proper measurement instrument for their research question.

Title

A simulation study of repeated covariate equating

Author(s)

Michaela Vařejková[1,2], Patrícia Martinková[1,2], Eva Potužníková[2]

[1]Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic; [2]Charles University, Prague, Czech Republic

Abstract

When performing test equating with non-equivalent groups and without an anchor test, one potential solution how to adjust for group differences is to substitute the anchor test score with covariates, such as grades or scores from another test. A key assumption for this approach is that the conditional distribution of test scores, given the covariates, is the same in all groups. In this work, we conduct a simulation study to investigate how different types of violations of the same conditional distribution assumption can affect the resulting equated scores. We consider two non-equivalent groups differing in ability. As covariates, we use two binary variables (which can, for example, refer to student's educational status and school type) and one continuous variable referring to a score from another test. The continuous variable was generated to be correlated with the other two binary variables. In the simulations, we explore several scenarios differing in sample size and the relationship between the test score and the continuous covariate. We show that if the assumption of equal conditional distribution is not met due to the fact that the covariate itself is measured using different test forms, the accuracy of the resulting equated scores can be improved by equating the covariate before incorporating it into the primary test scores equating algorithm.

Title

The mediating effects of gambling motives between depression, anxiety, and financial self-efficacy in problematic gambling.

Author(s)

Laura Maldonado-Murciano[1,2,3], Zsolt Horvárth[4,1], Andrea Czakó[1,4], Zsolt Demetrovics[1,4], Belle Gavriel-Fried[5]

[1]Center of Excelence in Responsible Gaming, University of Gibraltar, Gibraltar, Gibraltar; [2]Faculty of Psychology, University of Barcelona, Barcelona, Spain; [3]Institute of Neurosciences, University of Barcelona, Barcelona, Spain; [4]Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary; [5]Bob Shapell School of Social Work, Tel Aviv University, Tel Aviv-Yafo, Israel

Abstract

Aim: Problematic gambling has been related to depressive and anxiety symptoms and low financial self-efficacy (FSE).The present study aimed to examine the mediating effect of gambling motives in this relationship. Methods: The sample included 421 individuals (27.79% women, meanage=39.29). They were recruited through gambling treatment centres, Gamblers Anonymous groups, a social-media campaign, and land-based lottery and sport-betting kiosks in 22 cities in Israel between July and September 2022. The eligibility criteria were: having Israeli citizenship, being aged over 18, and having gambled at least twice in the past month. The eligible criteria for treatment centres and GA groups were individuals who were in treatment in the past year. The Problem Gambling Severity Index (PGSI), the Patient Health Questionnaire for Depression and Anxiety (PHQ4) scale, the Financial self-efficacy scale (FSES), and a modified version of the Gambling Motives Questionnaire (GMQ-R-14) were applied to assess health-related and gambling-related characteristics. Findings: A structural equation analysis with structural equation modeling showed that PG is directly positively significantly influenced by depression (standardised effect = .20, $p<.001$) and directly negatively significant effect of the FSE (standardised effect = .20, $p<.001$). These predictors also indirectly affected PG through escapism motivation (standardise effect = .46, $p<.001$). The model shows that escapism is the only gambling motive which mediates the relationship between depression and anxiety and FSE and PG. Conclusions: the escapism motive for gambling is the only motivation type which mediates between depression and anxiety and FSE and PG,

Title

Psychometric analysis of the University Entrance Examinations in Spain

Author(s)

Alejandro Veas[1], Elena Govorova[2], José-Antonio López-Pina[1]

[1]University of Murcia, Murcia, Spain; [2]University of Oviedo, Oviedo, Spain

Abstract

University Entrance Examinations (Spanish acronym: EBAU) are a key assessment tool for access to tertiary education. In Spain, as happens in other countries, they are based on examination standards of mandatory and modality subjects that have been studied during the previous course. In quantitative research, given the possible factors associated with grading, there have been several attempts in Europe to improve objective grading criteria. In this context, special attention has been addressed to inter-subject comparability using a variety of statistical procedure. Considering the Construct Comparability Approach as the theoretical framework applied in the United Kingdom, this study aims to implement the Partial Credit Rasch model for analyzing inter-subject comparability in the Spanish EBAU calls from 2017 to 2021.

As main results, the unidimensionality of the models were confirmed, with appropriate fit indices. The different subjects had an adequate difficulty in comparison with students' ability parameters. Score category did not have an adequate fit, specially in the lowest scores. Variations in subject difficulties between autonomous communities across time are also explored.

As main conclusions, and according to previous research, Rasch analysis constitute one of the main statistical strategy to implement inter-subject comparability approach. Implications about the use of standards and score criteria are discussed, together with general recommendations to ensure quality measurement.

Title

A Causal View on Bias in Missing Data Imputation: The Impact of Problematic Auxiliary Variables on the Norming of Test Scores

Author(s)

Erik Sengewald[1], Katinka Hardt[2], Marie-Ann Sengewald[3]

[1]German Federal Employment Agency, Nuernberg, Germany; [2]German Armed Forces, Köln, Germany; [3]Leibniz Institute for Educational Trajectories, Bamberg, Germany

Abstract

Among the most important merits of modern missing data techniques such as multiple imputation (MI) and full-information maximum likelihood (FIML) estimation is the possibility to include additional information about the missingness process via auxiliary variables. Our presentation emphasizes the importance of carefully selecting auxiliary variables, instead of using as many variables as available in the imputation model. First, the selection of auxiliary variables is outlined from the perspective of causal theory and framed in the context of norming. Then, we provide a proof-of-concept simulation study to investigate the potentially biasing effect of certain auxiliary variables (i.e., instrumental variables and colliders). In addition, we investigate an empirical norming example that uses ability test data from the German Federal Employment Agency. In this individual diagnostic setting, missing data can occur by design and according to certain missingness predictors, as the ability tests are administered under specific situations. For investigating a representative group, missing test scores can be imputed, but problematic auxiliary variables can amplify or even induce bias. This is the case in our empirical study, in which we compare norms obtained from imputed data of an individual compilation (IC) with norms from a standardized compilation (SC) of tests that ensures complete data. The imputation model uses different sets of auxiliary variables and allows for discussing the impact of problematic predictors in practice.

Title

The Diagnostic Potential of Process Data from a Computer-Based Simulated Supermarket

Author(s)

Philine Drake[1], Johannes Hartig[1], Manuel Froitzheim[2], Gunnar Mau[3], Hanna Schramm-Klein[2]

[1]DIPF | Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany; [2]University of Siegen, Siegen, Germany; [3]DHGS German University of Health and Sports, Berlin, Germany

Abstract

While the measurement and modeling of competencies typically focuses on task outcomes, behavioral differences during task completion are often not considered. With digital technologies, competence assessments can provide process data as additional information about the skills and strategies of test takers. Funded by the German Research Foundation (DFG) we focus on the so-called purchasing literacy of children and investigate elementary school children's self-control as an important aspect of their purchasing literacy in a simulated supermarket. To this end, 130 children were asked to shop on a limited budget and work through a given shopping list. We processed the data of this task in three ways: First, we combined process and product data into a common partial credit score for a differentiated assessment of task performance. Second, we derived theory-based behavioral indicators from the log data. By means of a structural equation model, we confirmed that the covariance between them could be explained by a factor of self-control. Within the structural equation model, we also investigated whether self-controlled behavior mediated the relationship between self-reported impulsivity and task performance. This could not be confirmed, even though self-controlled behavior was positively related to task performance. Third, using dynamic time warping and cluster analyses, we explored which patterns of change could be observed for the indicators of self-control during task processing. With the successful theory-driven extraction of construct relevant facets based on log data from a computer-based task, our study demonstrates the diagnostic potential inherent in such data.

Title

`Bifactor`, a flexible and fast R package for exploratory structural equation modeling of hierarchical structures

Author(s)

Marcos Jiménez[1], Francisco Abad[1], Eduardo García-Garzón[2], Luis Eduardo Garrido[3], Vithor Franco[4]

[1]Universidad Autónoma de Madrid, Madrid, Spain; [2]Universidad Camilo José Cela, Madrid, Spain; [3]Pontificia Universidad Católica Madre y Maestra, Santiago de los Caballeros, Dominican Republic; [4]Universidade São Francisco, São Paulo, Brazil

Abstract

Exploratory structural equation modeling (ESEM) is becoming increasingly popular because it avoids the misspecification problems of its confirmatory counterpart. However, flexible statistical software for dealing with hierarchical structures within the ESEM framework is lacking. Consequently, ESEM applications in fields like personality and intelligence, where traits are usually embedded within general factors, cannot be adequately implemented. To surpass this limitation, we developed the `bifactor` R package. `bifactor` can fit exploratory factor models with correlated general factors that are orthogonal to the specific factors while maintaining desirable features from the ESEM framework such as the possibility of correlating errors and conducting multigroup analyses. Moreover, all the fitting algorithms implemented in `bifactor` are built upon C++ code, making it well-suited for rotating large factor patterns in a short time. A description of all the features available in the `bifactor` package is offered.

Title

Enhancing Psychometrics with Interactive ShinyItemAnalysis Modules

Author(s)

Jan Netík[1,2], Patrícia Martinková[1,2]

[1]Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic; [2]Charles University, Prague, Czech Republic

Abstract

ShinyItemAnalysis is an R package and Shiny application widely used to teach psychometric concepts and conduct psychometric analyses without any coding experience (Martinková & Hladká, forthcoming). In this work, we present a new feature that has been introduced in the latest version of ShinyItemAnalysis, called "SIA modules". These modules, written completely in R using the Shiny package, allow researchers and practitioners to offer new analytical methods for wider use.

SIA modules are designed to integrate with and build upon the ShinyItemAnalysis app (Martinková & Hladká, 2018), enabling them to leverage the existing infrastructure for tasks such as data uploading and processing. They can access a range of outputs from various analyses, including item response theory models, exploratory factor analysis, or differential item functioning models. Because SIA modules come in R packages (or extend the existing ones), they may come bundled with their own datasets, use compiled code, etc.

We provide two SIA modules for demonstration: A module for estimation of inter-rater reliability in grant proposal peer-reviews under range restriction (Erosheva et al., 2021), and a didactic showcase of computerized adaptive testing (CAT) that covers the main CAT ideas and demonstrates the possibility to use the module's own data as well as the data uploaded by the user.

In summary, the SIA modules offer an innovative and interactive way for psychometricians to share their research and advances in methodology. This feature makes it easier for researchers and practitioners to explore psychometric concepts and analyses in a user-friendly way.

Title

Additional Results on the Performance of Location-Scale Models in Meta-Analysis: A Simulation Study

Author(s)

Desirée Blázquez-Rincón[1], Wolfgang Viechtbauer[2], José Antonio López-López[3]

[1]Universidad a Distancia de Madrid, Madrid, Spain; [2]Maastricht University, Maastricht, Netherlands; [3]Universidad de Murcia, Murcia, Spain

Abstract

Location-scale models are a useful tool in the field of meta-analysis since they allow the influence of moderator variables on the mean (location) and variance (scale) of the distribution of true effects to be studied at the same time. The implementation of location-scale models for meta-analysis was recently added to the metafor package for the R statistical software. In previous conference presentations, the results of a Monte Carlo simulation study comparing different estimation methods (maximum or restricted-maximum likelihood estimation), significance tests (Wald-type or likelihood-ratio tests), and methods for constructing confidence intervals for the scale coefficients (Wald-type and profile-likelihood intervals) were presented in terms of rejection rates and coverage probabilities. However, due to time constraints, other important results were not offered. With the present work, our goal is to present additional results regarding the bias and mean squared error, in the first place, of the estimates of the scale coefficients and, secondly, of the heterogeneity values for the levels of the moderator variable. Results are discussed with respect to the estimation method, the type of moderator variable (dichotomous or continuous), the heterogeneity values given to the levels of the moderator variable, the number of studies within the meta-analysis, and the average sample size of the included studies.

Title

Reproducibility and Data Sharing in Meta-analyses on the Effectiveness of Psychological Interventions

Author(s)

Rubén López-Nicolás[1], Daniël Lakens[2], José Antonio López-López[1], María Rubio-Aparicio[1], Alejandro Sandoval-Lentisco[1], Carmen López-Ibáñez[1], Desirée Blázquez-Rincón[1], Julio Sánchez-Meca[1]

[1]University of Murcia, Murcia, Spain; [2]Eindhoven University of Technology, Eindhoven, Netherlands

Abstract

In recent years concerns on the credibility of psychological research have emerged. Reproducibility of scientific results could be considered as the minimal threshold of it. In this study, our purpose was to assess the reproducibility of a set of published meta-analyses. From a random sample of 100 papers containing at least one meta-analysis on the effectiveness of interventions in psychology, 217 meta-analyses were selected. We first tried to retrieve the original data by recovering a data file, recoding the data from document files (pdf, doc), or on request. Second, through a multi-stage workflow, we tried to reproduce the main results of each meta-analysis. The original data were retrieved for 146 meta-analyses. Of these, in the first stage 52 showed a discrepancy larger than 5% in the main results, in 25 of them, this discrepancy was solved with minor adjustments, or correction of coding errors. In the remaining 27, different issues were identified in an in-depth review of the papers, such as reporting inconsistencies, lack of some data, or transcription errors. Current practices of data sharing in meta-analyses hamper the reusability of meta-analytic data. In addition, the implementation of new tools would help to avoid certain errors in the meta-analysis reporting process.
Funding: Project financed by the Region of Murcia (Spain) through the Regional Program for the Promotion of Scientific and Technical Research of Excellence (Action Plan 2022) of the Seneca Foundation - Science and Technology Agency of the Region of Murcia (grant no. 22064/PI/22). Grant PID2019-104080GB-I00 funded by MCIN/AEI/ 10.13039/501100011033.

Title

A Reliability Generalization Meta-analysis of the Fear of COVID-19 Scale (FCV-19S)

Author(s)

Desirée Blázquez-Rincón[1], Raimundo Aguayo-Estremera[2], Zainab Alimoradi[3], Elahe Jafari[3], Amir Pakpour[4]

[1]Universidad a Distancia de Madrid, Madrid, Spain; [2]Universidad Complutense de Madrid, Madrid, Spain; [3]Qazvin University of Medical Sciences, Qazvin, Iran, Islamic Republic of; [4]Jönköping University, Jönköping, Sweden

Abstract

The widespread administration and multiple validations of the Fear of Covid-19 Scale (FCV-19S) in different languages have highlighted the controversy over its underlying structure and the resulting reliability index. In the present study, a meta-analysis based on structural equation modeling (MASEM) was conducted to assess the internal structure of the 7-item, 5-point Likert-type FCV-19S version, estimate an overall reliability index from the underlying model that best reflected the internal structure (one -equivalent factor, one congeneric factor, or two-factor models), and perform moderator analyses for the model-implied inter-item correlations and estimated factor loadings. A Pearson inter-item correlation matrix was obtained for 48 independent studies, from which a pooled matrix was calculated following a random-effects multivariate meta-analysis. The results from the One-Stage MASEM analysis showed that the two-factor model properly fitted the pooled matrix, while the -equivalent and congeneric one-factor models did not. Nevertheless, the use of a bifactor model exhibited the predominance of the general factor over the domain-specific ones. High omega coefficients were obtained for the entire scale (.91) and the psychological (.83) and physiological (.83) symptoms subscales. Moderator analyses evidenced an increase in the estimated factor loadings, as well as in the reliability of the FCV-19S, when the standard deviation of the total scores increased and when the FCV-19S was administered to specific (vs. general) populations. The FCV-19S can be therefore considered as a highly related two-factor scale whose reliability makes it suitable for clinical and research purposes.

Title

A Reliability Generalization Meta-analysis of the Spanish Burnout Inventory (SBI)

Author(s)

Raimundo Aguayo-Estremera[1], Pedro Gil-Monte[2], Desirée Blázquez-Rincón[3]

[1]Universidad Complutense de Madrid, Madrid, Spain; [2]Universitad de Valencia, Valencia, Spain; [3]Universidad a Distancia de Madrid, Madrid, Spain

Abstract

Burnout is a response to chronic work stress that occurs when the individual feels overwhelmed and powerless to cope with difficulties at work, that affects a wide range of professional fields. Epidemiological data concerning this syndrome reflect the seriousness of the problem. Those who suffer from the syndrome usually show health problems of a psychosomatic, emotional, attitudinal, and behavioral nature. In addition, burnout has negative effects on organizations (sick leave, reduced performance) and for the users of the service. The Spanish Burnout Inventory (SBI) is a 20-item 5-point Liker-type scale that allows a rapid assessment of the prevalence of this regarding four dimensions: enthusiasm toward the job, psychological exhaustion, indolence, and guilt. The extensive administration of the SBI gives us the opportunity to synthesize and support the conclusions regarding some of its psychometric properties with samples of teachers from a wide variety of countries. The goal of the present work is to study the reliability of the SBI based on the model that better fits its internal structure. To do so, the One-Stage meta-analytic structural equation modeling (MASEM) technique is applied to the empirical Pearson correlation matrices collected in 32 independent studies. First, a combined correlation matrix is obtained following a fixed-effect and random-effects model to study the evidence of heterogeneity. Then, the four-factor model is fitted to the combined matrix. Based on the goodness of fit indices and the heterogeneity evidence, a reliability index is computed according to the four-factor structure and potential moderator variables are examined.

Title

Dimensionality assessment with categorical variables: an approach based on bootstrap

Author(s)

Francisco J. Abad[1], Luís E. Garrido[2], Marcos Jimenez[1], Rodrigo S. kreitchmann[3], Miguel A. Sorrel[1]

[1]Universidad Autónoma de Madrid, Madrid, Spain; [2]Pontificia Universidad Católica Madre y Maestra, Santiago de los Caballeros, Dominican Republic; [3]IE Universidad, Madrid, Spain

Abstract

Dimensionality evaluation is a critical step in factor analysis, as selecting the appropriate method for this process is essential to ensure accurate and reliable results. In this study, we investigated the effectiveness of two methods for evaluating dimensionality with ordinal variables: parallel analysis (PA) and a new method based on the unbiased root mean square residual (uSRMR), estimated by bootstrap. We manipulated several conditions, including sample size, number of indicators, factor loading size, factor correlation, and error population, to investigate the accuracy of these methods across different conditions. We also considered the accuracy in the recovery of the uncertainty in the estimated number of factors. Our findings revealed that both PA and uSRMR were reliable in evaluating dimensionality across various conditions. Overall, this study provides valuable insights into the effective application of PA and uSRMR in evaluating dimensionality when using factor analysis with ordinal variables. Our findings suggest that researchers should carefully consider the specific conditions of their data when selecting an appropriate method for dimensionality evaluation. Also, the use of the uSRMR method with bootstrap can provide an informative approach for evaluating dimensionality in factor analysis with ordinal variables, particularly in recovering the uncertainty in the estimated number of factors.

Title

The Model of Individual Methodological Orientations: Operationalization in the Linguistic Measurement System.

Author(s)

Sławomir Pasikowski, Martyna Jarota

University of Lodz, Łódz, Poland

Abstract

The aim of the poster is to present the issue of individual methodological orientations and propose a model of quantity and a model of its measurement. In this regard, the theoretical background of the introduced category showing its semantic location in the network of closely related concepts is outlined, in order to then derive a theoretical construct and a model using the category of oppositions occurring in the methodological discourse, based on the theory of orientation codes by Kazimierz Obuchowski and James J. Gibson's concept of affordances. This model allows to capture methodological orientations in a transversal way in relation to the divisions and classifications operating in the discourse of social science methodology while creating the possibility of building individual characteristics. The model was then operationalized using Maria Nowakowska's linguistic measurement system, which is a development of Lotfi Zadeh's fuzzy measurement system based on fuzzy set logic. The result was a complex model whose logical consistency was confirmed in the course of the verification procedure.

Title

To Caesar what is Caesar's? The responsibility of the experimenter in the first place!

Author(s)

María Paula Fernández[1], Guillermo Vallejo[1], Pablo Livácic-Rojas[2], Ellián Tuero-Herrero[1], Feliciano Ordóñez[3,4]

[1]Faculty of Psychology, Oviedo, Spain; [2]Santiago de Chile University, Chile, Chile; [3]Alfonso X el Sabio University, Madrid, Spain; [4]Padre Ossó Faculty (Affiliated Center University of Oviedo), Oviedo, Spain

Abstract

One of the biggest problems that every researcher faces is the possibility of losing data throughout the entire research process. Even if the investigation is cross-sectional, even when the data is collected in one day, and in a small period of that day, we are going to assume that in one hour, data loss is possible, and in that case, it is unlikely that data loss is completely random. In that case, data loss will most likely be random. The skill and foresight of the investigator can place a reasonably small limit on such data loss. On other occasions, the research is also cross-sectional, because only one pre-treatment measure (of several dependent variables or control variables), one post-treatment measure, and perhaps one or two follow-up measures are recorded. The data loss may be random, in part, but it is likely significant non-random data loss, and the culprit is most likely the researcher or experimenter. How many scientific articles does the researcher acknowledge that he is responsible for data loss? Review research is carried out in five different areas of knowledge and this aspect is evaluated. The results are indisputable. The recognition of having some responsibility for the loss of data must be residual because we have not yet found representation.

Title

Shhhhhhhh, be careful,..., there it is possible to lose data!!!

Author(s)

María Paula Fernández[1], Jose Antonio Labra[1,2], María Dolores Seijo[3], Francisca Fariña[4], Virgínia Daniela da Silva[5]

[1]Faculty of Psychology, Oviedo, Spain; [2]Cantabria University, Santander, Spain; [3]Faculty of Psychology, Santiago de Compostela, Spain; [4]Faculty of Education and Sport Sciences, Vigo, Spain; [5]School of Psychology, Minho, Portugal

Abstract

Data loss is a living, dynamic, opportunistic, self-serving, and threatening threat throughout the investigation process. It is democratic. Data loss occurs in basic, applied, cross-sectional, longitudinal research, in the field of medicine, psychology, engineering, and any scientific field (in non-scientific and para-scientific fields as well). In this investigation three completely different investigations are shown. In one of them, a comparative causal investigation on parenting and co-parenting is carried out. In another investigation, a complex case-control investigation is carried out to assess the convergent validity of classical cognitive tests with a battery of functional tasks to detect mild cognitive impairment. In a third investigation, the adaptation to the population of adolescents and pre-adolescents of a questionnaire on body image is carried out, and later a comparative causal investigation is carried out. An open protocol was developed to avoid losing data throughout the process of each of the investigations. As a result, the percentage of data loss was less than 5%. In conclusion. It is possible to reduce the data loss to a bearable level so that the statistical procedures developed to deal with this problem are effective.

Title

Loss of data, heterogeneity, variables that need to be controlled for, and various correlated dependent variables. A solution through multiple Imputation.

Author(s)

Guillermo Vallejo[1], María Paula Fernández[1], Pablo Livácic-Rojas[2]

[1]Faculty of Psychology, Oviedo, Spain; [2]Santiago de Chile University, Chile, Chile

Abstract

It is usual in applied research, and in basic research as well, to test the effect of more than one independent variable, and generally, these variables are manipulated by the researcher keeping them perfectly crossed. It is also common to need to keep some scale variable, or more than one, under control, and it is generally done through statistical control. In the course of the research, data is frequently lost, to a greater extent in the control group, but the experimental group is not exempt from this. As a consequence, the groups are unbalanced, and asymmetrically empty of data, and the researcher has to face an induced heterogeneity and an opportunistic loss of data. This research shows the development of the formulation to implement the multiple imputation procedure in the MANCOVA and shows the result of using this analysis through a Monte Carlo simulation experiment. It is concluded that the proposed procedure can be used with a guarantee since the validity of the statistical conclusion is protected.

Title

Educational assessment as classification: cluster analytical procedures in large scale studies

Author(s)

Gediminas Merkys, Sigitas Vaitkevičius, Daiva Bubeliene, Leonidas Sakalauskas

Vytautas Magnus University, Kaunas, Lithuania

Abstract

It is planned to share methodological experiences gained during a large-scale study. It is about the heuristic capabilities of K-Means cluster analysis discovered during a total census study in Lithuania. The data (N=340,000) cover educational achievement from the 4th to the 12th grades. Additive indices – z-scores and percentiles were created from the individual test scores. A model of 6 clusters - types of educational achievement - was obtained by clustering mathematical-scientific and humanities skills. By clustering repeated measures, a model of 9 clusters reflecting statistical types of educational trajectories was obtained. Both cluster patterns discovered are very stable in all grades and age groups of students. The distance between the extreme groups of achievement on the z-scale is about 3 standard deviations in both cluster models. In contrast to the latent class analysis, the K-Means analysis has no goodness of fit parameters. Here, some circumstances become fundamental. Are the clusters theoretically interpretable? Is it possible to validate individual clusters and the entire model using principles of construct and criterion validation? It is theoretically predicted that clusters should be specific with respect to the gender of the student, various social variables, etc. If the system of such hypotheses is confirmed, a positive decision is made about the validity of a particular cluster classification model. The analysis of statistically homogeneous types makes it possible to identify various external variables affecting educational achievement more precisely than is possible by using traditional causal analyses and working with mixed samples that are not typologically differentiated.

Title

Bayesian Sample Size Determination for Multilevel Models with Longitudinal Data

Author(s)

Ulrich Lösener, Mirjam Moerbeek, Herbert Hoijtink

Utrecht University, Utrecht, Netherlands

Abstract

A priori sample size determination is essential in designing trials in a cost-efficient manner and in avoiding underpowered or overpowered studies. Also, reporting a solid justification for a certain sample size forces the researcher to think about key aspects of their study such as hypotheses, design, and statistical model. Most often sample size calculations are based on null hypothesis significance testing (NHST), an approach that has recently received severe criticism. As an alternative Bayesian evaluation of informative hypotheses has been developed. Informative hypotheses can be formulated based on researchers' theoretical and/or empirical expectations and can include order restrictions of multiple estimands. Bayes factors are used to quantify the relative support in the data for a certain informative hypothesis without suffering from some of the flaws present in NHST. Sample size calculations in this framework rely on simulations and have only been studied recently. Available software for this is limited to simpler models such as ANOVA and t-test, in which independence of observations is a crucial assumption. However, this assumption is rendered untenable when employing a longitudinal design where observations are nested within individuals. In that case a multilevel model should be used. This paper aims to provide researchers with a tool to perform sample size calculations for multilevel models with longitudinal data in a Bayesian framework. To this end, I discuss the results of a simulation study for various realistic scenarios and introduce an open source R function that enables researchers to tailor the simulation to their specific situation.

Title

Sample Size Determination for Cluster Randomized Trials Using Bayes Factor

Author(s)

Camila Natalia Barragan Ibañez, Mirjam Moerbeek, Herbert Hoijtink

Utrecht University, Utrecht, Netherlands

Abstract

The cluster randomized trial design is extensively used in social, behavioural, and biomedical sciences. Complete groups, such as schools or families, are randomized to treatment conditions. For instance, the researchers can assign a school or class to a condition to evaluate the effectiveness of educational programs. Traditionally, the sample size for this type of design is based on null hypothesis statistical testing using p-values. This approach uses the effect size, Type I error, and statistical power. However, focusing on p-values has led to questionable research practices and difficulties interpreting results. Testing hypotheses using the Bayes factor can overcome the disadvantages of using p-values. The Bayes factor quantifies the relative support for each hypothesis considered in the study. Previous research has proposed methods to determine the sample size, but these are limited to the t-test, Bayesian Welch's test, and ANOVA. Building on Fu (2022), the present study aims to determine the sample size necessary to evaluate informative hypotheses using the Bayes factor in cluster randomized trials. A simulation study is performed for cluster randomized trials with a one-period parallel-group design. The Bayes factor is calculated using the R package bain, and it is determined whether the criterion for sample size is met. The results of the simulation study will be presented, and general recommendations for a priori sample size determination in cluster randomized trials will be discussed.

Title

Evaluating Estimation Quality in Multilevel Random Effects Models

Author(s)

Denny Kerkhoff[1], Fridtjof W. Nussbeck[2]

[1]Bielefeld University, Bielefeld, Germany; [2]Konstanz University, Konstanz, Germany

Abstract

Random effects models for the analysis of nested data on three levels place high demands on the sample sizes, i.e., the number of level-3 clusters, the number of level-2 subclusters per cluster, and the number of level-1 units per subcluster. To evaluate the estimation quality of parameters and standard errors in relation to sample sizes and model specification, indicators such as relative and absolute parameter estimation bias measures, statistical power, and coverage rates are commonly computed through Monte Carlo simulations. Interpretation of these indicators should be handled with care due to their interrelatedness. For example, power and coverage rates are influenced by estimation bias, and there might be a trade-off between (relative) bias and variance in point estimates. I present findings on the relative and absolute bias, power, and coverage rates in three-level random effects models for the empty model, a random slopes model (Kerkhoff & Nussbeck, 2019, 2022), and contextual analysis models with manifest and latent predictors, which can be used to derive sampling strategies to ensure sound estimation quality. Particular attention is paid to the meaning and applicability of the estimation quality indicators, possible alternative measures, and comparison to the use of estimation quality measures in related fields.

Kerkhoff, D., & Nussbeck, F. W. (2019). The influence of sample size on parameter estimates in three-level random-effects models. Frontiers in Psychology, 10, 1067. https://dx.doi.org/10.3389/fpsyg.2019.01067 [javascript%3A]

Kerkhoff, D., & Nussbeck, F. W. (2022). Obtaining sound intraclass correlation and variance estimates in three-level models: The role of sampling-strategies. Methodology, 18(1), 5–23. https://doi.org/10.5964/meth.7265 [javascript%3A]

Title

Challenges to designing degree research projects in the Peruvian context, recovering methodological foundations.

Author(s)

Maria Mercedes Henriquez-de-Urdaneta

Universidad de Piura, Piura, Peru

Abstract

The study aims to describe the methodological challenges to designing degree research projects in the Business Administration Program, where recovering the role of conceptualization, operationalization, methodological complementarity, and mentoring was crucial in ensuring the research construct to the Peruvian context as a developing country where the enterprises are primarily informal. The work is based on the theoretical postulates of Pichler and Reiter (2023), Brown and Lindsay (2015), Polo (2006), Rao and Reddy (2013), Sellés (2013), Nubiola (2010), among others. During the experience, two contextualizations were added; the first consisted that Perú wasn't mandatory to develop and present a research project as a degree requirement, which implied planning basic research training for the academic, administrative staff, and students, identifying the need of design instruments to verify research topic knowledge as bases for the dimensions and indicators selection, methodology selection, items drafting assertiveness, and project tests assignations. The second attention was the Peruvian situation during the pandemic due to the country was one with the most extensive mandate to dictate and develop all academic activities 100% online for two years, appearing the challenge of redesigning methodological frameworks, where operationalizing was crucial to adapt the design and integrated other methods. Supporting the entire process, the methodological complementarity and mixed methods as adaptation tools for the contexts considered. Moreover, the experience success allows assisted more than 400 undergraduate research degree projects, generating the call to replicate it in masters' programs, like Engineering and Marriage and family with interdisciplinary projects.

Title

Conditional Standard Errors of Measurement for personality tests in personnel selection: An empirical comparison of IRT and Generalizability Theory approaches

Author(s)

Rene Gempp[1], Sergio Valenzuela[2]

[1]Universidad Diego Portales, Santiago, Chile; [2]Universidad Católica de Chile, Santiago, Chile

Abstract

Although the use of Conditional Standard Errors of Measurement (CSEM) to quantify the uncertainty of test scores is advocated by psychometricians, their use is infrequent in practical testing applications such as personality assessment of personnel selection, likely because most personality tests have been developed within the factor analytic and Classical Test Theory (CTT) traditions, while the gold standard methods for estimating CSEM are based on Item Response Theory (IRT). Additionally, a few simple methods for estimating CSEM within the CTT paradigm are valid for dichotomous items rather than the polytomous items typically found in personality questionnaires. This study empirically compares two methods for estimating CSEM in a personality test with polytomous items: an IRT-based method and a simplified version of the Generalizability Theory (GT) approach suggested by Brennan (1998, 2001). Both procedures were applied to a personality test of 44 items, each with five response options, based on the five-factor model, answered by 949 job applicants. In GT, a one-facet model was applied to each of the five dimensions, and the relative CSEM were estimated using the simplified procedure developed by Brennan (1998, 2001). In IRT, each dimension was calibrated using Samejima's Graded Response Model, and the CSEM were estimated using the method described by Wang, Kolen, and Harris (2000). The results show that the CSEM obtained using the GT approach are comparable to those estimated using IRT. The practical implications for measurement practitioners are discussed.

Title

Measuring and controlling prestige in vocational interests through a technique adapted from Peabody's quadruples

Author(s)

Leonardo Mose, Felipe Valentini, Livia Mendes, Bruno Santos, Isadora Francesco

Universidade São Francisco, Campinas, Brazil

Abstract

Occupational prestige is a critical aspect of vocational interest measurement. However, both constructs' variances are confounded, making it difficult to discriminate between them correctly. We could remove this confusion by adapting Peabody's technique (originally used for assessing social desirability), balancing items measuring the same interest in different prestige poles. Therefore, this study aimed to estimate and control prestige in vocational interests. For that, we developed the Prestige Assessment Inventory in Vocational Interests (PAIVI) using Peabody's adapted technique. The final version of PAIVI consisted of 274 items and 45 facets of basic interests, assessing the six RIASEC types. Each facet of the instrument was set by four items, two of which were developed to assess high-prestige activities and two items assessing low-prestige activities. A sample of 848 adults aged 18 or over answered the PAIVI and the Personal Globe Inventory (PGI), which also measures occupational prestige. We analyzed the data using confirmatory factor analysis and fitted a bifactor model to orthogonalize the prestige from the content of interests. The model fitted the data well. The results showed that it is possible to extract a general prestige factor. Several general factor loadings were estimated in the theoretically expected direction (positive loadings for high-prestige activities and negative loadings for low-prestige activities). We suggest that orthogonalizing the prestige variance from vocational interests makes it possible to assess interests with less influence from occupational prestige.

Title

Does acquiescence also impact the factor structure of vocational interest inventories?

Author(s)

Leonardo Mose, Felipe Valentini, Jennifer Bathaus, Giselle Machado, Lorena Queiroz, Giovanna Leopoldino, Maria Theotonio, Gustavo Martins

Universidade São Francisco, Campinas, Brazil

Abstract

In vocational interest inventories, it is common for a portion of scores variance to be explained by a general factor. However, whether the general interest factor consists of a substantial construct or a response bias, such as acquiescence, is not yet known. For controlling acquiescence, it is essential to use balanced questionnaires in which each item has an opposite semantic pair. In this sense, this work aimed to construct a Balanced Inventory of Vocational Interests (BIVI) to estimate and control acquiescence in vocational interests. A total of 204 positive and 204 negative items were developed based on a Holland dimensional model (RIASEC). We used the following strategy to construct the balanced scale: the positive item should be written by starting the sentence with: "It would be interesting..." or "I would like to...". Example: "I would like/It would be interesting to drive a tractor." On the Other hand, the negative opposite pair should be elaborated starting with: "It would be boring..." or "I would hate it...". After content analysis by experts, 150 items were chosen to be applied in an empirical study. The sample consisted of 452 college students from a Brazilian university in São Paulo. We analyzed the 150 items using exploratory factor analysis. The results showed that the acquiescence variance was extremely low. Even after controlling for bias using ipsatization, a general factor was estimated using Schmid-Leiman orthogonalization. This could suggest that the general dimension of interests is unrelated to the acquiescent response style.

Title

Time-varying effects in psychometric survey data

Author(s)

Jinghui Liang, Dale Barr

School of Psychology and Neuroscience, University of Glasgow, Glasgow, United Kingdom

Abstract

Psychometric analyses often apply factor analysis or structural equation modelling to estimate a psychological scale's reliability and validity. Yet these approaches usually ignore the effects caused by item context, sequence, and other "human factors", thereby violating statistical assumptions. Since these effects unfold over time, the sequence of measurements is likely to reflect a mixture of the underlying construct and time-varying effects such as fatigue or practice effects. To investigate this possibility, we developed an open-source platform for online survey data collection that collects participants' ratings along with reaction times for each item in a measurement. This platform also allows managing presentation orders to comprise fixed, fully randomized, and quasi-randomized sequences in order to examine whether time-varying effects can be counteracted across participants via balanced experimental design strategies. With this empirical dataset, we conducted linear mixed-effect model (LMEM) and generalized additive mixed models (GAMMs) analyses to examine whether time-varying fluctuations were present. We found that participants have idiosyncratic patterns in their ratings and reaction speeds which are likely to be overlooked when looking at overall means. Based on these results, the future direction of this project will be refining factor analysis methods and adjusting autocorrelated residual terms in factor analysis to generate a psychometric modelling method that can disentangle psychological constructs from time-structured noise in measurement.