11th Scientific Meeting of the SIS Group
"Statistics for the Evaluation and Quality in Services"

BOOK OF SHORT PAPERS

Editors
Andrea Bucci
Alfredo Cartone
Adelia Evangelista
Andrea Marletta

IES PESCARA 2023

STATISTICAL METHODS FOR EVALUATION AND QUALITY: TECHNIQUES, TECHNOLOGIES AND TRENDS (T^3)

EDIZIONI Il Viandante

A project evaluation study on multiset Likert scale data

Uno studio di valutazione su dati multiset in scala Likert

Violetta Simonacci, Marina Marino, Maria Gabriella Grassia and Michele Gallo

Abstract This work is part of the evaluation proposal for the experimental phase of the ClassMate Robot project, promoted by the Protom Group. The experimentation consists in testing how a newly developed AI device for social education is received in a classroom environment. To assess usability, likability, and social impact pre- and post-trial surveys were administered to the participating students of 4 schools. The data is arranged in multi-block architectures and then summarized with IRT tools. A classic non-parametric approach is employed for testing before and after differences. Post-experimentation results are explored via PARAFAC2 to model school differences while accounting for a multiset structure.

Abstract *Questo lavoro fa parte della proposta di valutazione per la fase sperimentale del progetto ClassMate Robot, promosso dal Protom group. La sperimentazione consiste nel testare come un nuovo dispositivo AI per l'istruzione viene percepito in un contesto scolastico. Per valutarne usabilità, likability e impatto sociale sono stati somministrati questionari pre e post agli studenti partecipanti di 4 scuole. I dati sono inseriti in architetture multiblocco e riassunti con strumenti IRT. Un approccio non parametrico viene utilizzato per testare le differenze prima-dopo. I risultati post-sperimentazione saranno esplorati via PARAFAC2 per modellare le differenze tra scuole tenendo conto della struttura mulitset.*

Key words: AI for education, impact assessment, IRT, PARAFAC2, survey data

Violetta Simonacci
Dept. of Social Science, University "Federico II", Naples, Italy, e-mail: violetta.simonacci@unina.it

Marina Marino
Dept. of Social Science, University "Federico II", Naples, Italy, e-mail: marina.marino@unina.it

Maria Gabriella Grassia
Dept. of Social Science, University "Federico II", Naples, Italy, e-mail: mgrassia@unina.it

Michele Gallo
Dept. of Human and Social Sciences, University "L'Orientale", Naples, Italy e-mail: mgallo@unior.it

1 Project description and introduction

Protom Group S.p.A. is the first Italian Knowledge & Technology-Intensive (KTI) company with a cutting-edge profile in the field of digital transformation. In 2021, they set in motion a pioneering business project based on technology for education, known as ClassMate Robot (CMR). The idea behind CMR is to use Artificial Intelligence (AI), by introducing an in-house built social robot archetype, to bring upon the conventional Italian school framework innovative teaching and learning processes.

The project carried out through Protom Robotics and Scuolab includes the collaboration with the Projects of Intelligent Robotics and Advanced Cognitive System (PRISCA) Lab of the University of Naples "Federico II" for the development of the software infrastructure and the scientific support of the Department of Social Sciences (DiSS) of the University of Naples "Federico II". In detail, DiSS played an active role in defining software requirements, outlining the educational framework, and implementing an experimental phase in the real context of 4 Italian schools (Junior High and High School level). DiSS is also responsible for final reporting and for carrying out a full assessment study. The experimental phase will officially close at the end of the 2022-2023 school year.

Impact assessment includes a handful of qualitative and quantitative tools. Interviews and case studies are accompanied by the development of business and social intelligence paths. A detailed plan for data harvest during the experimentation was devised together with Protom Group, specifically imagining the potential use of collected information.The goal of the quantitative assessment is not only to measure the success of the project but also to provide useful tools for the improvement of the AI device and the joint cloud platform.

The plan includes two types of data collection tools: automatic detection via device and platform; and the administration of surveys to the entire cohort of students. Questionnaires are to be submitted at two time points to allow a comparison between the final perception and the initial expectation. In the specific context of this presentation, we focus on survey data only for brevity reasons.

In detail, two data-sets are considered for each participating school: 1) Survey at t_0, student responses to the initial questionnaire, nested by school, which include 28 items on a four-point Likert scale (grouped in 4 thematic blocks) and sociodemographic items; and 2) Survey at t_1, student responses to the final questionnaire, nested by school, encompassing 8 blocks of items with the same 28 questions of the previous survey and an additional 33 items all on a four-point Likert scale. A more detailed description of the data is provided in Section 2.

The aim of this work is to implement an initial exploratory study of survey data by keeping in mind two major data characteristics: the ordinal nature of Likert scale items and the presence of a nested design, as school grouping is likely to have an impact on student responses. The analytical goal is to answer several questions concerning:

1. Survey validity: can we obtain valid summary measures for each thematic block?

2. Pre- and post-experimentation differences: are there significant changes in student responses on common blocks before and after experimentation?
3. Measure interactions: how do the detected constructs interact with each other?
4. Nesting effects: What is the imprint of school grouping?
5. Role of External variables: Are there sociodemographic group differences?

The methodological design developed to address these queries is articulated in the following manner. To start, summary measures are built for all blocks using an Item Response Theory (IRT) approach in order to properly treat Likert scale items and assess survey validity. The impact of experimentation on the 4 blocks common to Survey at t_0 and Survey at t_1 is verified via Wilcoxon signed-rank tests. To study measure interaction in the nested samples at t_1, an exploratory perspective was preferred, using the PARAFAC2 model. Section 3 outlines the methodology in more detail. Section 4 presents the project outlook.

2 Survey Data

To properly assess CMR performance it was deemed necessary to quantify the likability of the device and its validity as a teaching tool through the administration of a post-experimentation survey (at t_1). Questions were specifically developed to measure: students' general perception of the CMR (8 items on usability and likability), students' comfort level using the CMR (9 items), students' perception of CMR impact on school results (5 items), students' perception of platform likability (6 items).

To gather more information on students and classroom environment a set of general questions were also added to measure school well-being, following [8]. These Likert-scale items are divided into 3 blocks: relationship with teachers (7 items on trust, support, recognition), relationship with classmates (6 items on acceptance, trust, friendships), and sense of self-efficacy (10 items). In addition, a collection of socio-biographical information (8 questions on gender, parental education and employment, living situation, and grade) and a block on the relationship with technology on a Likert scale (10 items) were added. It was decided to submit well-being questions and the relationship with the technology block also before the experimentation (at t_0) to test the CMR social impact. All Likert scale items are on a 4-point system where the options were "1 = NOT AT ALL", "2 = A LITTLE", "3 = ENOUGH", "4 = A LOT".

Four classes, located in four different schools in Italy, were selected for the trial (Rome, Carrù, Dalmine, and Verona). A total of 96 students participated in the project.

Collected data can be easily organized in nested data structures. The responses to Survey at t_0 can be arranged in a multi-level object $\mathscr{X}^{t_0}$ subdivided in school-by-item-block tables. Formally, for the k-th school and s-th item block, we have a generic table $\mathbf{X}_{ks}^{t_0}$ holding the scores of the I_k students of school k on the J_s items of the block s as follows:

$$\mathbf{X}_{kb}^{t_0} = \begin{bmatrix} x_{1_k 1_s} & \cdots & x_{1_k j_s} & \cdots & x_{1_k J_s} \\ \vdots & \ddots & & & \vdots \\ x_{i_k 1_s} & & x_{i_k j_s} & & x_{i_k J_s} \\ \vdots & & & \ddots & \vdots \\ x_{I_k 1_s} & \cdots & x_{I_k j_s} & \cdots & x_{I_k J_s} \end{bmatrix} \tag{1}$$

where $k = 1, \ldots, 4$ and $s = 1, \ldots, 4$ for a total of 16 tables. If the tables a juxtaposed we have a 96 rows by 28 columns object:

$$\mathscr{X}^{t_0} = \begin{bmatrix} \mathbf{X}_{11}^{t_0} & \cdots & \mathbf{X}_{1s}^{t_0} & \cdots & \mathbf{X}_{1S}^{t_0} \\ \vdots & \ddots & & & \vdots \\ \mathbf{X}_{k1}^{t_0} & & \mathbf{X}_{ks}^{t_0} & & \mathbf{X}_{kS}^{t_0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{X}_{K1}^{t_0} & \cdots & \mathbf{X}_{Ks}^{t_0} & \cdots & \mathbf{X}_{KS}^{t_0} \end{bmatrix} \tag{2}$$

Similarly, for Survey at t_1, a multilevel object $\mathscr{X}^{t_1}$ can be built. The object includes 32 school-by-item-block tables, as in the matrices 1 and 2, where the only difference is that $s = 1, \ldots, 8$.

An object $\mathscr{G}$, nested by school can also be built for student socio-demographic variables, in which student information is collected for the 8 described items.

3 Methodology

The methodological flow of this work can be summarized in the following phases:

STEP 1: Each item block in $\mathscr{X}^{t_0}$ and $\mathscr{X}^{t_1}$ is tested for consistency and reliability throughout schools using Cronbach's alpha and the Automated Item Selection Procedure (AISP) [6]. Problematic items may be considered as separate measures or sub-blocks may be formed if diagnostics suggest modification.

STEP 2: For each consistent block a measurement scale is identified throughout schools which represents a one-dimensional latent trait. To attain an interval scale, an IRT approach is used, which is based on the probabilistic relation between item difficulty and subject ability. In detail, the Partial Credit Rasch Model (PCM) [10] will be employed. The model can be described as follows:

$$\log(P_{ilc}/P_{il(c-1)}) = U_i - V_l - F_{lc} \tag{3}$$

Here the probability P_{ilc} for the i-th subject of responding in the category c rather than in the category $(c-1)$ in reference to the l-th item is a function of the subject ability U_i, the item difficulty V_l and the rating scale structure F_{lc}. This model yields simplified versions of $\mathscr{X}^{t_0}$ and $\mathscr{X}^{t_1}$ where item blocks are replaced by summary measures. We obtain reduced-size objects, only nested by school, which can be defined as the multiset tensors $\underline{\mathbf{T}}^{t_0}(I_k \times N \times K)$ and $\underline{\mathbf{T}}^{t_1}(I_k \times M \times K)$, with generic

element $t_{i_k nk}$ and $t_{i_k mk}$ respectively. They represent a collection of the K tables $\mathbf{T}_k^{t_0}(I_k \times N)$ and $\mathbf{T}_k^{t_1}(I_k \times M)$, where $n = 1,\ldots,N$ and $n_1 = 1,\ldots,M$ indicate the set of summary measured obtained at t_0 and t_1.

STEP 3: Paired Wilcoxon signed-rank tests are performed on all measures to assess if significant differences were recorded at t_1 on the well-being and technology relationship items. A non-parametric test was preferred due to the sample size.

STEP 4: To study measure interaction, $\underline{\mathbf{T}}^{t_1}$ is decomposed via PARAFAC2. The PARAFAC2 model [5, 7] can be described as a less restrictive version of standard PARAFAC which can also be applied when the data tensor is not fully-crossed (presenting a dimension of varying size).

Ordinary PARAFAC [1, 4] decomposes a fully-crossed tensor $\underline{\mathbf{T}}(I \times M \times K)$ based on the parallel proportional profiles principle [2] and the assumption of complete trilinearity. In detail, it assumes that the obtained latent variables correspond to real constructs which hold proportional patterns throughout levels. As a result, it yields only three loadings matrices $\mathbf{A}(I \times R)$, $\mathbf{B}(M \times R)$ and $\mathbf{C}(K \times R)$, where R is the number of extracted factors, one for each dimension of the tensor, in the following manner:

$$\mathbf{T}_k = \hat{\mathbf{T}}_k + \mathbf{E}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^t + \mathbf{E}_k \quad k = 1,\cdots,K. \tag{4}$$

Here $\mathbf{T}_k$ is the generic frontal slices of $\underline{\mathbf{T}}$, i.e. an $(I \times M)$ matrix for a given occasion k. $\mathbf{D}_k$ is a diagonal matrices holding the kth row of the third-mode loading matrix $\mathbf{C}$ and lastly $\mathbf{E}_k$ is the frontal slice of the residual tensor $\underline{\mathbf{E}}$. The model is unique under mild conditions.

PARAFAC2 relaxes the assumption of trilinearity by allowing for different loading matrices across levels in one of the three dimensions (conventionally the first dimension). This is particularly useful in the case of multiset data, where there are incomparable observation units across samples. The model can thus be adjusted as follows:

$$\mathbf{T}_k = \hat{\mathbf{T}}_k + \mathbf{E}_k = \mathbf{A}_k\mathbf{D}_k\mathbf{B}^t + \mathbf{E}_k \quad k = 1,\cdots,K \tag{5}$$

The main difference is that the model generates K loading matrices $\mathbf{A}_k$. To ensure the uniqueness advantage also to the PARAFAC2 model, a restriction is imposed that the loading matrices $\mathbf{A}_k$ only differ in terms of rotation, i.e. the cross-product (covariance or correlation) matrix $\mathbf{A}_k{}^t\mathbf{A}_k$ is constant over k.

STEP 5: PARAFAC2 results are visually studied also for assessing the behavior of different socio-demographic groups with special attention to gender differences and parental education level.

4 Project outlook

The experimental phase officially terminates in May 2023. Nonetheless, most surveys have already been collected. A first, though incomplete, analysis has been implemented which demonstrates the feasibility and effectiveness of the planned

methodology described in this paper. Complete results will be available for discussion during the presentation.

Some methodological advancements are also being considered. The advantage of using a multilevel approach for the extraction of summary measures, following [3], rather than standard PCM, will be explored. Alternative methods to build summary measures will also be implemented and analyzed in a comparative fashion, such as the approach proposed in [9], where a multivariate compositional analysis is carried out to extract bipolar constructs known as log-contrasts. Lastly, in the second stage of the evaluation process, Data Analytics will be studied in detail to also evaluate technical performance and its impact on likability and overall student experience.

References

1. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika 35(3):283–319 (1970)
2. Cattell, R.B.: "Parallel proportional profiles" and other principles for determining the choice of factors by rotation. Psychometrika 9(4):267-283 (1944)
3. Doran, H., Bates, D., Bliese, P., Dowling, M.: Estimating the multilevel Rasch model: With the lme4 package. Journal of Statistical software 20:1-18 (2007)
4. Harshman, R.A.: Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis. UCLA Working Papers in Phonetics 16:1–84 (1970)
5. Harshman, R.A.: PARAFAC2: Mathematical and technical notes. UCLA Working Papers in Phonetics, 22:30-44 (1972)
6. Hemker, B.T., Sijtsma, K., Molenaar, I.W.: Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken I RT model. Applied Psychological Measurement 19(4):337-352 (1995)
7. Kiers, H.A., Ten Berge, J.M., Bro, R.: PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model. Journal of Chemometrics: A Journal of the Chemometrics Society 13(3-4):275-294 (1999)
8. Marzocchi, G.M., Tobia, V.: QBS 8-13: Questionari per la valutazione del benessere scolastico e identificazione dei fattori di rischio. Edizioni Centro Studi Erickson (2015)
9. Simonacci, V., Gallo, M.: Statistical tools for student evaluation of academic educational quality. Quality & Quantity 51: 565-579 (2017)
10. Wright, B.D., Masters, G.N.: Rating scale analysis. MESA press (1982)