
Editorial

Elena C. Papanastasiou*

School of Education,
University of Nicosia,
46 Makedonitissas Avenue, Nicosia 2417, Cyprus
Email: Papanastasiou.e@unic.ac.cy
*Corresponding author

Hanna Eklöf

Department of Applied Educational Science,
Umeå University,
901 87 Umeå, Sweden
Email: hanna.eklof@umu.se

Biographical notes: Elena C. Papanastasiou is the Dean of the School of Education of the University of Nicosia. She has received her PhD in Measurement and Quantitative Methods from Michigan State University and Honours BSc in Elementary and Kindergarten Education from The Pennsylvania State University. She has held academic positions at the University of Kansas, University of Cyprus as well as University of Nicosia. Her current research interests involve the topics of test development, psychometrics, and quantitative research methods. She currently represents Cyprus in the General Assembly of the International Association for the Evaluation of Educational Achievement (IEA), while in 2014 she was elected in the Standing Committee of the IEA. In 2018, she was awarded the status of Fellow of the Association of Educational Assessment-Europe.

Hanna Eklöf is an Associate Professor at the Department of Applied Educational Science, Umeå University, Sweden. She has received her PhD in Educational Measurement from Umeå University, and her research interests include the psychology of test-taking, test-taking behaviour and motivation, large-scale testing, and measurement quality issues. She is currently the Principal Investigator of a four-year project funded by the Swedish Research Council, exploring student effort and test-taking behaviour through process data in PISA, and co-investigator in other externally funded projects investigating math anxiety, motivation, working memory and performance. She also teaches in the area of educational assessment, and is member of the steering group for the national research school Quantitative Research Methods in Education.

Researchers, educators and policy makers know rather little about how students and other ‘classroom stakeholders’ actually interact with tests, since in most cases they only rely on examinees’ final item or test scores. Recent technical and methodological advancements in the area of large-scale testing have provided us with opportunities to better understand the testing process. This can be achieved through the modelling of student test-taking behaviours, test-taking strategies, as well as through their interactions with non-cognitive

variables and performance. Such developments have the potential to be used for evaluating psychometric measures, as well as for obtaining further evidence on the degree of validity of test scores. As a result, the possibilities in analysing process data, primarily obtained by computer-generated log data, have lately attracted great interest in quantitative educational research. The hopes attached to these types of data are high, but the research area is yet not very well developed. So, through this special issue, we aim to provide some insights on the potential of process data in the field of assessment.

Process data, primarily obtained from log files from computer-based tests and questionnaires provide a rich array of non-traditional data sources that can be used for describing the interactions that take place between test items and examinees. Such data can include time-stamped events showing the time needed for answering items, keystroke events, mouse-clicks, eye-tracking information, etc. But handwritten notes or traces of student behaviour from paper-and-pencil tests (e.g., the crossing out of answers) can also be used for the same purposes. Thus, the examination of such data in combination with the final answers that have been provided by examinees have the potential to contribute to methodological developments in this area, and to provide further insights into our understanding of the testing process.

The purpose of this special issue and the set of papers included are to showcase ways in which the underexplored area of process data, beyond the mere use of right or wrong responses, can be utilised to provide additional information about examinees, items and test quality. This is done through the dissemination of four empirical studies, performed in different assessment settings, using different approaches, and exploring different research questions, all however related to 'what goes on behind the scenes' of assessment. The four papers individually and together provide rich input into ways of analysing process data, which may increase our understanding of the process of test-taking, the meaning of test results, while adding to the validity of our interpretation and use of test scores. Two of the papers engage with the analysis of computer-generated traces of interactions between tests and test-takers (log files) in relation to different kinds of feedback during test-taking. The other two papers make use of paper-based tests and investigate patterns of non-responses and aberrant test behaviour, respectively. Each paper provides valuable suggestions as to how we may discover patterns hidden underneath the overt assessment and the assessment results.

How do students react to and act upon immediate feedback on their performance during assessment? The first paper: 'Using cluster analysis to explore students' interactions with automated feedback in an online earth science task', by Zhu et al. explores this question by using event-log data from an online earth science course module to discover patterns in high-school students' activities when writing and revising scientific arguments. The use of an automated scoring engine provided immediate feedback to students and the purpose of the analysis was to identify how students interacted with the automated scoring and feedback given to improve students' revision of these arguments. Cluster analysis was applied to data with the goal of identifying subgroups of students who exhibited similar review and revision behaviours. Findings indeed revealed distinct clusters, characterised by different revision patterns, initial item scores and item score changes. The study contributes both by demonstrating cluster analysis as a viable method for analysing log file data in the form of action sequences, and by revealing how students may react to and interact with immediate feedback, while examining how different groups of students display varied patterns in their interactions.

How do instructors manipulate settings, and how do students behave under different settings in a computer-based assignment? The second paper: 'Hints, multiple attempts, and learning outcomes in a computer-based formative assessment system', by Choi and Bogucki focuses on formative assessment in a computer-based system featuring assignments targeting content in a college-level physics book. Similar to the first paper by Zhu et al., the interest can be said to be on the effects of feedback during test-taking, but in this case, on learner actions in relation to instructor actions who control the assessment settings (e.g., turning features on/off, thereby allowing them to be more or less lenient and formative). Raw log data were mined out of an online platform database and from these, authors defined measures such as hint usage, performance, persistence, and conscientiousness. Findings from linear and logistic regression suggest that instructors' use of settings that encourage hint opening and multiple attempts was associated with learners' higher persistence and conscientiousness, but with lower performance. Empirically this study could show that different settings were associated with different behaviours, a finding that can be useful in shaping instructions on how to provide optimal assessment settings. Methodologically, it shows that mining process data from a digital platform can result in defining meaningful measures that in turn can be used to make inferences about learners' learning behaviour and outcomes.

What can non-responses tell us that responses cannot? The third paper in this volume: 'Can non-responses speak louder than words? Examining patterns of item non-response in TIMSS 2015', by Papanastasiou uses data from a large-scale international comparative study, TIMSS 2015 (paper-based assessment, Grade 4 students), to explore non-responses as indicative of how students approach the test. In the paper, it is rightly argued that although missing data is dealt with all the time in research, it is rarely treated and investigated as possibly carrying important information about how students engage with the test per se. Therefore, the purpose of this study was to identify possible variations in patterns of non-responses and their predictors based on item and student characteristics. A series of linear regression analyses revealed that differences do exist between countries in terms of tendency to omit items, even among countries having similar levels of overall achievement. In addition, certain student and item characteristics were also found to be related to non-response patterns, though patterns varied between countries. Thus, findings suggest an interplay between student characteristics, item characteristics and non-responses, indicating that non-responses cannot be considered as missing at random but rather indicative of differences in test-taking behaviour and strategies. Although the reasons why students in some countries tend to omit items to a higher degree than students in other countries need further exploration, it is suggested that clearer guidance to all students on how to approach test-taking could be important.

How can we determine if a school's performance on a high-stakes test is aberrant, and possibly indicative of cheating? While the third paper by Papanastasiou deals with a low-stakes assessment (TIMSS) where students may be less motivated to place effort on the test and partly therefore choose to omit items, the fourth paper: 'Screening for aberrant school performances in high-stakes assessments using influential analysis' by Christiansen et al. plays out in a high-stakes context where motives for (spuriously) raising performance may be present. Aberrant performance is often difficult to prove, but here a method is proposed which allows screening for aberrant school performances in high-stakes assessments. The authors developed a technique that allows incorporating longitudinal school measurements with the same test on different cohorts with the

purpose of identifying atypical (aberrant) results at the school level. More specifically, the study used influential analysis under a Bayesian approach, modelling achievement levels with a beta inflated mean regression model and measuring divergence to determine aberrancy. Although the study is more on the technical side, it has a very applied meaning and findings seem promising. Detecting aberrant school performances is as difficult as it is important in true high-stakes assessments where incentives for raising performance are high. Therefore, this paper demonstrates both in simulation and in the context of a Peruvian high-stakes national assessment that the proposed method seems to be able to identify ‘unexpected results’. Using techniques like these could have important practical implications in the future.

The four papers included in this special issue clearly show that the analysis of process data can be very useful in different assessment settings in order to gain insight on different features of assessment and assessment behaviour. Together they span an impressive range of assessment contexts – low-stakes and high-stakes, large-scale and small-scale, national and international, summative and formative, etc. They also employ different methods to explore their research questions. However, they have one thing in common: they use information that is available ‘around’ the key variables of the assessment contexts, and by doing that, they inform us more about the how’s and why’s of the outcomes in terms of assessment results that we can observe. Each paper contributes to the research on process data by demonstrating ways to make use of existing but neglected information, ways of developing meaningful variables from data mining procedures or employing methods that are suitable for increasing our understanding of data and findings.

In quantitative educational research, we currently experience a turn towards increased interest in what we can learn about students through analysing their interactions with assessment. We probably still have a long way to go along this line of research, but all of the above papers are good ambassadors for how this research could be pursued. All papers represent promising lines of research, substantially and/or methodologically and together, they surely display the potential of process data in the field of assessment.