



**The Psychometrics Centre  
Department of Psychology  
School of the Biological Sciences  
University of Cambridge**

**Examining the Gender Differences in Mathematical  
Problem-Solving Performance on the Croatian  
Secondary School Final Examinations**

**Essay  
Application for the Voya Kondic Memorial Prize**

Josip Šabić, PhD candidate  
Department of Psychology  
Faculty of Humanities and Social Sciences  
University of Zagreb

My name is Josip Šabić and I work in the National Centre for External Evaluation of Education (NCEEE) in Zagreb as the Head of the Psychometrics Unit. As a psychometrician, reviewer of educational test items, standard setting coordinator and teacher trainer, I am involved in nearly all phases of educational test construction. Most of these tests, such as Croatian secondary school final examinations, are high-risk and high-stakes tests and require the highest level of validity and reliability.

In 2010, I started my PhD in Psychology at the University of Zagreb, where my key modules have been Educational Psychology, Psychometrics and Cognitive Psychology. A year ago, I could only dream of spending one part of my doctoral studies in one of the world's best and most famous universities. This changed with the help of the British Scholarship Trust and because of it, I have been a visiting student in the Psychometrics Centre at the Department of Psychology in the School of the Biological Sciences, University of Cambridge from November 1, 2013 to January 31, 2014. I was the British Scholarship Trust scholar in November of 2013 and this stipend was crucial in making my stay in Cambridge possible.

The Psychometrics Centre is one of the world's leading institutions in the area of assessment and it gathers brilliant young scholars from the whole world. My supervisor was Professor John Rust, Director of the Psychometrics Centre, Director of Research in the Department of Psychology and also one of the leading authorities in the field of testing and assessment. During November, I worked on a project called "Examining the Gender Differences in Mathematical Problem-Solving Performance on the Croatian Secondary School Final Examinations", which was closely related to my doctoral research at the Department of Psychology in the University of Zagreb. In this period, I worked hard on conducting statistical and psychometric analyses of data for my doctoral research. I also started writing my thesis; both the introductory theoretical part of the thesis and the interpretation of the obtained results.

The University of Cambridge and the School of the Biological Sciences provided me with excellent facilities to carry out my research. I was allowed to use a very well equipped library at the Department of Psychology. The University of Cambridge has access to a vast number of licensed online collections of full text journal articles, which was crucial for elaborating the theoretical part of my dissertation, since most of the important research papers in the field of psychometrics and educational psychology are published in the form of articles. Studying literature from this field helped me to better understand new and very complex methodologies which I used in my research, such as differential item functioning (DIF) and differential bundle

functioning (DBF), as well as deepen my knowledge about patterns of gender differences in mathematics examinations that were obtained in previous studies. During the work on my analyses, I had the privilege to discuss my problems and dilemmas with Professor Rust and other experts from the Psychometrics Centre who used DIF methodology in their former studies and who are more experienced than me in conducting complex statistical procedures.

The taxonomy of content and cognitive skills expected to produce gender differences in mathematics (Gallagher et al., 2000), which was used in my research, was based on the results of studies that investigated gender impact. The use of DIF and DBF methodology gives a new insight into this area. Banks (2013) found only 16 studies in which authors used DBF methodology. In 14 studies authors used DBF in exploring educational test results, which points to the conclusion that these methodologies are still insufficiently used in the educational context.

The objective of my study was to explore the taxonomy of content and cognitive skills expected to produce gender differences in mathematics (Gallagher et al., 2000). The most striking result of this study is gender gap on items that require verbal skills. The direction of difference in favour of males was not expected on these items. The gender gap on these items is relatively large even after controlling for total test result. *Controlling for total result* means that results on items that require verbal skills are compared for males and females with similar total test results. In other words, on this group of items females have lower results than males who have the same level of Mathematics knowledge. This result is very important information for secondary school Mathematics educational practice in Croatia because teachers can now help females to bridge the difference.

From the methodological and statistical point of view, probably the most interesting result is the next one: if males have a higher average result than females on items with some characteristic, this difference will be smaller after controlling for total result; if females have a higher average result than males on items with some characteristic, this difference will be higher after controlling for total result. This result probably occurred because males had slightly higher average results in both types of schools and both years. This pattern is found for every item characteristic from both years and for both types of schools in this study, except for one.

In January 2014 I held a talk about my research titled “Differential Item and Bundle Functioning in Educational Testing Contexts”, as a part of the Cambridge Psychometrics Centre Seminars. This was a huge honour for a doctoral student like me and a wonderful opportunity to

present my work to other experts from the University of Cambridge. My research and doctoral thesis have benefited greatly from the discussion with my colleagues after this talk.

During the visit, I also had time to participate in other activities of the Department of Psychology. For example, I attended a course “Basic and Advanced Statistical Techniques Using R”, which was held by Dr. Aleksandr Kogan. This was a great opportunity for me to improve my skills in R, which is a powerful free software environment for statistical computing and graphics. In R, it is possible to perform a wider range of statistical analyses than with some of the commercially available statistical computer programs. Also, I deepened my knowledge of linear multiple regression, which is one of the widely used techniques in educational psychology. For instance, by using this methodology it is possible to better understand the relationship between different variables (e.g. socio-economic status, intelligence, gender, personality traits, etc.) and educational test results.

I also attended a workshop on Item Response Theory (IRT), Computer Adaptive Testing (CAT) and the open source online adaptive test development platform “Concerto”, at which I learned how to develop online adaptive tests. In CAT, questions are selected in the test on the basis of the participant’s answers to previous questions for the purpose of maximizing the precision of the examination for that particular participant. In this procedure, the difficulty of the examination is readjusting to the level of ability of participant shown in previous items of the test. Every educational testing agency such as my agency NCEEE, should consider implementing this new methodology in its tests.

I also attended the Cambridge Psychometrics Centre Seminar “Multilevel Analysis: Application in Large Scale Assessments”, which was held by Prof. Dr. Igor G. Menezes, Federal University of Bahia, Brazil. I am planning to use this methodology in my future work at the NCEEE in Zagreb. Multilevel analyses are usually used in large scale assessments and could be implemented in the analyses of Croatian national examinations and secondary school final examinations which are carried out by the NCEEE. A nice feature of multilevel analyses is that they allow the researcher to use variables at different levels (e.g. student level, class level, school level, regional level, etc.) to explain the outcome variable (e.g. student test results). Ignoring relationships between variables from different levels can lead to wrong conclusions in one’s research, for example in predictions of student test results. As a short-time member of the University of Cambridge, I had an opportunity to obtain a copy of MLwiN software, created by

The Centre for Multilevel Modelling (CMM), which is based at the University of Bristol. In the near future, I am planning to attend the on-line course on the University of Bristol website about conducting multilevel analyses in this software and hopefully raise my level of expertise in this field.

After my return to Croatia, I am planning to organize a workshop about DIF and DBF analyses in statistical software SIBTEST for my colleagues from the Psychometrics Unit at the NCEEE and present them the results of my research. I am also planning to do a presentation for my colleagues in the doctoral studies at the Department of Psychology at the University of Zagreb. These results will also be presented to Mathematics examinations' item writers. Item writers are secondary school Mathematics teachers who are collaborating with NCEEE. After finishing my dissertation, I am planning to present the results of this research to secondary school Mathematics teachers. This is one of the first studies of educational test item bias in Croatia. It is important to sensitise teachers to be aware of the gender gap in some areas of secondary school mathematics with an emphasis on verbal problems that produce the largest gap.

It would be useful to implement methodologies used in this study in research on group differences in other school subjects such as mother tongue, foreign languages, etc. It is possible to use similar methodologies in measuring differences between other groups, i.e. between students from different geographical regions of Croatia. To enhance the fairness of educational tests and test validity, it is important not just to find items that are biased but also to identify the causes of item bias. It is important to understand why some groups of students have lower results on educational tests because results on these tests can have impact on students' future academic choices and career paths. Isolating item characteristics that differentially affect students from different groups is a crucial step before giving advice to teachers and other educational experts. Afterwards it is possible to make interventions in educational practice and teaching to help certain groups to overcome difficulties in specific parts of the curricula. If some item characteristic that is not important for the curriculum produces a gap between groups, it is possible to avoid that kind of items in future examinations.

In conclusion, I benefited from this scholarship and my stay at the Psychometrics Centre because I raised the level of my expertise in a large number of different modern statistical and psychometric procedures. The majority of these procedures is immediately applicable in analyses that are run on a daily basis in the NCEEE. Furthermore, at this moment, DIF and DBF

analyses are non-existent in Croatian psychometric practices. In the future I will have an opportunity to disseminate this knowledge to other Croatian psychometricians and test developers.

This experience has also improved my other skills. It has made me more competitive on the job market. During my stay in England, I also improved my English skills. I also learned a lot about other cultures and established many wonderful friendships with colleagues in the multi-ethnic and multi-national environment of the Psychometrics Centre. I am most grateful to the British Scholarship Trust for making this internship possible.

#### Literature

Banks, K. (2013). A Synthesis of the Peer-Reviewed Differential Bundle Functioning Research. *Educational Measurement: Issues and Practice*, 32(1), 43–55.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology*, 75, 165-190.