Topic Study Group No. 39: Large Scale Assessment and Testing in Mathematics Education

Rae Young Kim, Christine Suurtamm, Edward Silver, Stefan Ufer and Pauline Vos

Introduction

Topic Study Group 39 aimed to address issues related to large-scale assessment, evaluation and testing in mathematics at all levels. Sound large-scale assessment (LSA) has the potential to provide important feedback about students' mathematical thinking, about classroom mathematical culture, or about a country's curriculum emphasis. Furthermore, LSA can have a strong influence in mathematics education as it often defines the mathematics that is mediated, valued and worth knowing.

Our TSG sought contributions of research in and new perspectives on LSA in mathematics education. We saw these issues as falling into three main strands: purposes and use, design and development, and teacher-related issues. Prospective contributors were requested to address one or more of the following topics:

Purposes and Use

- Purposes and use of LSA in mathematics at the international, national, school, classroom, or individual level
- The use of assessment for learning, as learning, and of learning in mathematics as they relate to LSA

Co-chairs: Rae Young Kim, Christine Suurtamm.

Team members: Edward Silver, Stefan Ufer, Pauline Vos.

R.Y. Kim (✉)
Ewha Womans University, Seoul, Republic of South Korea
e-mail: kimrae@ewha.ac.kr

C. Suurtamm
University of Ottawa, Ottawa, Canada
e-mail: Christine.Suurtamm@uottawa.ca

G. Kaiser (ed.), *Proceedings of the 13th International Congress on Mathematical Education*, ICME-13 Monographs, DOI 10.1007/978-3-319-62597-3_66

- Policy issues such as how LSAs frame political discussions and decisions
- The communication and use of results from LSA in mathematics

Design and Development

- The development of LSAs which might include the conceptual foundations of such assessments
- Task design that values mathematical power including problem solving, modeling, and reasoning across disciplines, and that addresses the diversity of learners
- The design and implementation of alternative modes of LSA in mathematics (e.g., online, student investigations)

Teacher-related issues

- The design and development of LSA of teachers' mathematical and pedagogical content knowledge
- The impact of LSA on teachers' knowledge and practice

We initially received over 40 papers for the TSG covering a wide range of areas of interest from all over the world. We discussed how to organize the sessions and participated in reviewing the papers. Each paper was evaluated by two reviewers including co-chairs, team members, and the authors of the papers submitted to TSG 39. Based on the reviews of these papers, 12 of these contributions were chosen for extended papers, 14 were chosen for oral communication, and 12 were recommended for poster presentations. Considering the topics and issues of the papers, we categorized the papers into three extended paper sessions, three oral communication sessions, and one poster session (at general exhibition) facilitated by co-chairs and team members as chair. In addition, we had a joint session with *Topic Study Group 40: Classroom Assessment for Mathematics Learning* to share mutually interesting issues, ideas, and practices around assessment through intensive discussion. We collaboratively produced a pre-conference publication with the classroom assessment group as well. Since some papers were withdrawn, 11 papers were presented in extended paper sessions, 1 paper was presented in the joint session (along with 2 from TSG 40), 11 were presented in oral communication sessions, and 8 were shown in the poster session in the end.

All the sessions of TSG 39 were organized to create a sense of community among all the presenters and participants who share common interests and ideas about large-scale assessment to improve mathematics education. The participants contributed greatly to the sessions and brought in perspectives from a wide range of knowledge, experiences, and practices. They were asked to read all of the papers before coming to the TSG 39 sessions and to bring some questions and comments on the papers. We also generated online space to facilitate further discussion out of sessions. The following are the leading questions in the discussion:

- How do we ensure that we are assessing what is important to assess?
- What framework do people use in task design or assessment evaluation?

- What should be considered in task design?
- How do MKT items developed in one country transfer to other countries?
- What do we need to take into consideration when examining student achievement on LSA?
- In what ways can technology interact with assessment?
- How can LSA assessment be designed and used to improve student learning and equity?

Main Ideas and Discussions in Each Session

Each session consisted of three or four 15-min presentations, short questions and comments after each presentation, and a 20-min whole group discussion at the end. Although each session was originally organized by the main themes, various issues and questions related to several themes came up together in the sessions. Thus, we summarized what was presented and discussed by the main themes shown above: Purposes and use, design and development, and teacher-related issues.

Purposes and Use

More than 17 papers were presented regarding this main theme with various perspectives throughout the sessions. The presentations showed that large-scale assessments have been implemented for multiple purposes and uses in mathematics education. One group of papers focused on the use of large-scale assessments to evaluate systems and to make student placements. For instance, there are analyzing issues in specific regions such as gender and socioeconomic status (SES) in Brazil (e.g., Chagas and Kleinke) and the case of bonus points in Ireland (e.g., Treacy). Some papers presented the use of assessments to make student placements (e.g., Reddy) or to predict student performance by finding some factors or determinants (e.g., Alagoz and Ekici; Seifert, Eilerts, and Rinkens; Weitz and Venkat).

Another group of presentations showed that large-scale assessments could be used to reveal the features of student achievement and affective characteristics in certain contexts or across national contexts. Many papers focused on the analysis of student achievement in specific regions such as Taiwan (e.g., Tam and Leung), Belgium (e.g., Deprez, Nijlen, Ameel, and Janssen), and Thailand (e.g., Jaikla, Changsri, and Inprasitha) or across countries in terms of cognitive domains or levels (e.g., Kanageswari). While discussing several issues and concerns in each context, we also found commonalities across contexts.

The results from large-scale assessments contribute to analysis of factors related to student achievement. For instance, the relationship between self-efficacy and student achievement by their cognitive levels (e.g., Zhou, Liu, Q., and Liu, J.), the effects of socioeconomic status (SES) and opportunity to learn (OTL) at classroom and country levels on student achievement (e.g., Bokhove), the relationship

between the use of ICT and mathematics achievement (e.g., Kanoh), didactic contract (e.g., Ferretti, Gambini, and Giorgio), and the factors influencing affective characteristics (e.g., Hwang, Kim, H., and Kim, W.). Some papers suggested a natural model of analysis of student abilities (e.g., Dimitric) or items measuring students' geometric intuition (e.g., Bai, Huang, and Zhang). We discussed pedagogical and political issues from the results of studies as well as methodological concerns around data analysis and interpretation.

Design and Development

Many presentations brought up methodological issues around the design and development of tasks in large-scale assessments. For instance, the validity of the assessment (e.g., Bansilal; Grapin; Kasoka, Jakobsen, and Kazima), cross-cultural adaptations of measures (e.g., Marcinek and Patrová), cultural sensitivity and validity (e.g., Philpot), perceived task difficulty different from empirical one (e.g., Beitlich, Lehner, Strohmaier, and Reiss), and equivalent assessment design (e.g., Inekwe). In addition, many studies showed that individual or cultural differences in solving problems, especially word problem (e.g., Strohmaier, Beitlich, Lehner, and Reiss) or problems with realistic situations (e.g., Chen, Liu, Zhao, Song, and Li), could influence the reliability and validity of large-scale assessment.

Another group of presentations pointed out that large-scale assessments have often measured low level of cognitive demands (e.g., Dogbey and Dogbey; Drüke-Noe and Kühn), which could not reflect current goals in mathematics education. In order to enhance student learning through large-scale assessment, some presentations suggested new ways of evaluating student abilities by developing new items to measure geometric intuition (e.g., Bai, Huang, and Zhang) or providing a new guideline and prescription for interpreting problem situations with multicultural values (e.g., Djepaxhija, Vos, and Fuglestad).

Teacher-Related Issues

Although a relatively small number of papers focused on this theme, we discussed how the results from large-scale assessments could be used for improving teaching practice and teacher knowledge. Since teaching is a cultural activity in a situated context, we also discussed cross-cultural adaptation issues of using measures of Mathematical Knowledge for Teaching (MKT) from a certain context to another (e.g., Marcinek and Patrová) and considered qualitative approaches such as using video clips (e.g., Bruckmaier and Krauss).

We learned from the joint session with the classroom assessment group that large-scale assessment and classroom assessment could complement each other to improve mathematics teaching and learning. In particular, Burkhardt argued that high-stakes assessment could be "a tool for improvement" by playing the roles not only in assuring accountability of systems but also in "measuring student

performance", "defin(ing) performance goals for teaching and learning", and "largely determin(ing) the balance of classroom activities in most classrooms." This implies that large-scale assessment and classroom assessment can inform each other and enhance student learning in constructive ways.

Concluding Remarks

All the participants actively participated in the sessions and brought up interesting and important issues around large-scale assessments. We finally found that there were both decontextualized commonalities and contextualized differences across different contexts. In this sense, it was productive to collaborate with TSG 40, the classroom assessment group, to elaborate our discussions around assessments and improve assessments for student learning. We also came to the conclusion that further discussion needs to be continued to develop the emerging ideas from this topic study group.

Acknowledgements The contribution of all the authors and participants in TSG 39 are deeply acknowledged.