The Alberta Journal of Educational Research *Vol. XLV, No. 4, Winter 1999, 329-332*

W. Todd Rogers
University of Alberta

Introduction

Measurement and Evaluation: Current and Future Research Directions for the New Millennium

We are witnessing in Canada today, as in other countries, a marked increase in the use of tests and assessment. Large-scale achievement testing has become an important method for monitoring the quality of education in Canada. Concerned about the quality and effectiveness of their educational systems, the Council of Ministers of Education in Canada conduct national assessments in language arts, mathematics, and science. Further, nine of the 10 provinces have assessment programs at various grade levels. As well, Canada and five of the provinces participated in the 1993/94 Third International Mathematics and Science Study; data for Canada and five provinces were collected during 1999 as part of the Fourth International Mathematics and Science Study. Despite the apparent importance of these large-scale assessments, we still do not have a full understanding of the purposes and uses of these assessments and the contextual factors that need to be taken into account to better understand the performance of the students and the schools they attend.

At the classroom level, the day-to-day assessments of student learning are unquestionably one of the teacher's most demanding, complex, and important tasks. Every model of the teaching-learning process requires that teachers base their decisions—instructional, grading, and reporting—on some knowledge of student progress toward desired learning outcomes. Yet, given the prevalence of classroom assessment, our understanding of in-class assessment practices still needs to be more fully developed.

Eight articles are included in this special issue of the *Alberta Journal of Educational Research*. Together they address a sample of the issues that need to be addressed in the next millennium to ensure that the scores and information gained from large-scale assessments and from in-class teacher assessments can be reliably and validly interpreted (*Principles for Fair Student Assessment Practices for Education in Canada*, 1993). They were selected from a series of papers presented at the conference Measurement and Evaluation: Current and Future Research Directions for the New Millennium held in October 1998 in Banff, Alberta.

The first four articles in this special issue are devoted to large-scale assessments. In the first article, Bob Wilson raises questions about the validity of large-scale assessments, pointing out that the issues to be addressed are not

---

Todd Rogers is a professor in the Department of Educational Psychology and Director of the Centre for Research in Applied Measurement and Evaluation. He specializes in measurement and evaluation.

simple. He concludes his article with a set of thought-provoking questions that need to be addressed if large-scale assessment programs are to meet their varied goals.

The next three articles address validity issues like those identified by Bob. Over time we have seen the introduction of new, often complex techniques for conducting different aspects of a large-scale assessment. Large-scale testing agencies often feel compelled to adopt these new procedures. John Anderson suggests that any change should be "carefully evaluated to ensure that there is a high probability of improvement through change." While acknowledging the advantages of item response models (the new) for item banking, for example, he demonstrates that when producing estimates of student scores, the difference in the validity of interpretation between estimates determined using right-wrong scoring (the old) and estimated determined using item-response scoring were inconsequential.

Mark Gierl, Todd Rogers, and Don Klinger address issues of validity that arise when one test is translated into another language. An important issue in a bilingual country like Canada, unwanted systematic differences that are attributable to translation errors will serve to distort the inferences made from a score obtained from the translated test. Gierl et al. used both statistical and judgmental methods to identify the extent of differential item functioning (DIF) between English-speaking students who write large-scale achievement tests written in English and English-speaking French Immersion students who write the same achievement test translated into French. They conclude their article by calling for more research on methods for identifying DIF and on the need for researchers to focus on the actual cognitive processes used by the students as the students respond to items.

In the fourth article in this set, Philip Nagy and Randy Penfield explore ways of using large-scale assessment results at the individual classroom level. By means of a work in progress they introduce and examine some alternative, nonconventional, and simple-to-use methods for analyzing profile data to gain classroom and student diagnostic information.

The next three articles illustrate how cognitive psychology and psychometric measurement can be integrated to enhance the assessment of learning. Carl Frederiksen and Janet Donin provide a dynamic cognitive model to assess, authentically, student problem-solving in a computer-based coached learning environment. They conclude their article with a number of suggestions for future research on the application and extension of their model to incorporate, for example, diagnostic information, assessment of change, and use of the model in collaborative (as opposed to individual) learning situations.

Jackie Leighton, Todd Rogers, and Tom Maguire examine ways of modifying performance assessments in such a way that they can be objectively scored while at the same time calling for the *task* and *process* authenticity (Royer, Cisero, & Carlo, 1993) inherent in the item in its original form. The conjecture they tested was that when students score other students' work, they engage similar cognitive processes.

John Kirby provides a third way researchers are using cognitive psychology and measurement to increase the validity of the inferences drawn about student performance. He presents a theory of cognitive processing involved in

learning to read and demonstrates the degree to which measures derived from this theory are able to predict success in reading.

In the eighth article, Xin Ma illustrates the use of the recently developed hierarchical linear model (HLM) analysis (Byrk & Raudenbush, 1992). In this particular application, Xin used data from the Longitudinal Study of American Youth and a three-level model to examine differences in how males and females acquired mathematical skills during the secondary grades and to determine the relationship between their rates of growth and student- and school-level characteristics.

Final Words

The first seven articles in this special issue revolve around two central issues: increasing the validity and utility of large-scale testing and the combined use of cognitive psychology and measurement practice to enhance the validity of inferences we draw from our assessment instruments. Clearly more research into both issues is called for. Further, heeding the advice of Wilson and the call by Gierl, Rogers, and Klinger, what is also needed is a closer and firmer connection between large-scale testing and cognitive psychology. The eighth article illustrates the use of HLM, a statistical procedure that can be, and has been, used to assess the various contextual factors that need to be taken into account to better understand the performance of the students and the schools they attend (Willms, 1992).

Perusal of the titles of the papers presented at the conference reflects the need to examine further, for example, the reliability and validity of scores derived from performance assessments using structural equation modeling and generalizability theory, clarification of the nature of what is actually being assessed through the use of protocol analyses (Ericsson & Simon, 1993), clarification of the identification and nature of differential item performance, the policy implications of large-scale assessment programs, and the need to improve the assessment of attitudes and interests.

We must be careful, though, not to ignore the need for research into in-class, teacher-developed assessment practices. How teachers conduct assessments and how they unite the results of these assessments with their instructional decision-making and grading of their students need to be explored. Teachers, especially in the early grades, tend to place greater reliance on, and have more confidence in, their own judgments of student performance, but little is known about these kinds of activities.

A final item. The conference Measurement and Evaluation: Current and Future Research Directions for the New Millennium was held in honor of Tom Maguire, who retired from the University of Alberta in June 1997. A valued colleague, friend, and, for many, mentor, Tom's contributions to the development of measurement and the people working in the field of measurement in Canada and internationally were telling and valued. A true scholar, he has and will always have the respect of all of us. Tom, ***thank you!***

Acknowledgments

The conference on which this special issue is based was supported by the Social Sciences and Research Council of Canada, the Canadian Educational Research Association, the Faculty of Education at the University of Alberta, and Prentice-Hall, Inc. I would like to thank Mark Gierl and Rina Perez for their help in organizing the conference John Anderson, Mark Gierl, and Phil

Nagy who served as the reviewers for the articles submitted for publication in this special issue; and Mark Gierl for his helpful comments on this introduction.

References

Byrk, A.S., & Raudensush, S.W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis*. Cambridge, MA: MIT Press.

Principles for fair student assessment practices for education in Canada. (1993). Edmonton, AB: Joint Advisory Committee. (Mailing Address: Joint Advisory Committee, Centre for Research in Applied Measurement and Evaluation, 6-110 Education Centre North, University of Alberta, Edmonton AB T6G 2G5).

Royer, J.M., Cisero, C.A., & Carlo, M.S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63*, 201-243.

Willms, J.D. (1992). *Monitoring school performance: A guide for educators*. Washington, DC: Falmer.