Advances in Modern Assessment Models, Methods, and Practices
(Version: February 23, 2011)

Ronald K. Hambleton
University of Massachusetts Amherst, USA

I regret that I am not able to participate in the International Start-Up Conference on the new German Federal Ministry of Education and Research's Initiative, "Modeling and Measurement of Competencies in Higher Education." I have been working in the field of psychometric methods and educational assessment for over 44 years, and I believe this is the first time that I have failed to meet a commitment to speak at a regional, national, or international meeting. I had my passport stolen (and computer, telephone, glasses, and important papers too), and I was not able to obtain a replacement from the Canadian government in time to be in Berlin with you. Right until the last minute, I thought I would be successful, but I was not.

Had I been at the meeting, and with the 45 minutes I had been allocated to speak, I would have made the following comments:

I would have said that tremendous progress has been made in the field of assessment in the last 30 years or so: (1) we have seen the concept of criterion-referenced assessment (CRA) introduced for assessing competencies, and that concept is now well developed (for everything from defining the competencies to developing the assessments and validating them for use), and widely applied today in education, industry, and the military, (2) we have seen the transition from classical to modern testing theory, methods, and practices (perhaps the theory is better known as "item response theory" or 'IRT"), and the impact has been world-wide and highly significant, (3) we are seeing the expanded use of computers in assessment, and this use holds great promise, and (4) to a great extent because of the uses of computers in assessment, we are seeing a wide array of new item types for assessment being introduced (for example, these item types might include audio and video displays, be sequential, and often have candidates writing out responses that can be scored by the computer). I would have mentioned too about the pioneering work of one of your presenters, Professor Mark Wilson, and his current initiatives to define constructs, and develop new assessments by drawing on cognitive science, computers, and cognitively-based item response theory models. At the same time, many of these advances have not found much use to date in higher education. Clearly, there is lots of important work to be done and this point is a primary theme of the conference.

Next, I would have addressed each of the four advances.

With respect to criterion-referenced assessment (CRA), we need to recognize that this represents an approach to assessment where the student's performance is considered in relation to a well-defined domain of knowledge and skills or competencies (for, say, a university course). Performance standards are set and used to evaluate each student's

performance. The keys to success include (1) a well defined domain of knowledge and skills or competency statements to guide both instruction and assessment development, and (2) assessments that are closely matched to the domain of knowledge and skills, or competencies, that students are expected to master. CRAs were a response in the USA to norm-referenced assessments which were principally designed to provide normative information about students in relation to well-defined norm groups (e.g., all fourth graders in a region of interest). Norm-referenced assessments remain important, but they are not especially useful in the context of day-to-day instruction and assessment, or the evaluation of student performance in a university course, or competencies that might be expected of all university students, for example, in a country. Today, both norm-referenced and criterion-referenced assessments have important roles in education. Sometimes, NRAs are used in selection decisions and often CRAs are valuable to instructors in monitoring the progress of their students and providing course-end summative assessments.

Modern test theory was introduced by Georg Rasch from Denmark and Gerhard Fischer from Austria, along with Fred Lord and Allen Birnbaum from the United States in the 1950s and 1960s. For example, today, IRT plays a central role in TIMSS, OECD/PISA, and PIRLS, and many attendees may know of these testing programs but not about the central role of IRT. Today, too, the topic is being widely researched, developed, and applied around the world. IRT has several advantages over classical test theory (CTT) and methods because it provides an effective response, at least when models can be found to fit the available data, to some of the shortcomings of CTT: item statistics are examinee sample dependent, candidate scores are dependent on the particular choice of test items, the modeling is carried out at the test level, and more. The advantages of IRT include the features that there are many models available and approaches to assessing model fit are widely available, software is readily available to carry out analyses, item statistics are independent of the particular samples of students on which they are based. proficiency estimates do not depend on the particular choice of questions that are administered (e.g., this opens up the possibility for comparing students to each other or to performance standards without requiring all students to be administered the same test questions), unique estimates of imprecision for student scores are available, and much more. All is not perfect, but today, many testing agencies are making regular use of these IRT models. Had I been there, I would have shown how tests can be built and compared using item information functions and item efficiency curves, and potential bias can be studied. These are two of many popular applications.

The transition to computers for educational assessment represents another major trend. Computer-based testing (CBT) (1) permits more flexible scheduling of test administrations, (2) with automated scoring available, CBT means that reporting of test scores can be immediate, (3) new test designs can be put in place including adaptive tests which may reduce testing time to 50% with no loss in measurement precision, and (4) CBT provides the basis for new item types (to be considered next). Already in the USA many high stakes tests (such as year-end exams) are administered via a computer, and many times more, diagnostic tests, are now being administered via a computer to relieve some of the burdens of testing that are encountered by teachers. Of course new problems

are arising such as the need for large item banks, item inventory control, exposure controls, and more. At the same time, the potential for improving assessment practices via CBT is obvious and the volume of test administrations per day is increasing at a rapid rate. CBT seems particularly attractive in assessments for higher education students because all of the potential advantages would be appreciated by these students, more so, than students in elementary and secondary schools.

Finally, with the success so far of CBT, we are now seeing in the USA, the increasing development and use of new item types—item types that may involve audio and/or video stimuli, that may involve students manipulating graphs and tables on a computer screen, that require students to construct their answers to questions, and much more. Our first speaker at the conference, Professor Richard Shavelson, has been hugely important in developing new item types for assessing higher level cognitive skills in science, and the ETS representative on the program (Thomas Van Essen), is at a company that has been one of the leaders in developing new item types for computer-based assessments.

My own view is that all four of the developments I would have discussed—criterion-referenced assessments, item response theory and its applications, computer-based testing, and new item types, are going to make assessment of competencies in higher education more reliable and valid, and the assessments might be a lot more interesting for students, especially those aimed at higher level thinking skills, Good luck as you begin to build your research agenda, and ultimately implement new and improved assessments for students in higher education.

A Few Key References

Gierl, M. J., & Leighton, J. P. (Eds.). (2007). Cognitive diagnostic assessment for education: Theory and applications. New York: Cambridge University Press.

Hambleton, R. K. (2003). Criterion-referenced testing: Methods and procedures. In R. Fernandez-Ballesteros (Ed.), Encyclopedia of psychological assessment (pp. 280-283). London: Sage.

Hambleton, R.K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Thousand Oaks, CA: Sage Publications.

Mills, C. N., et al. (Eds.). (2002). Computer-based testing: Building the foundation for future assessment. Mahwah, NJ: Lawrence Erlbaum Publishers.

van de Linden, W. J., & Glas, C. A. W. (Eds.). (2009). Elements of adaptive testing. New York: Springer.