Rolf Strietholt
Wilfried Bos
Jan-Eric Gustafsson
Monica Rosén
(Eds.)

Educational Policy Evaluation

through International Comparative Assessments

WAXMANN

Rolf Strietholt
Wilfried Bos
Jan-Eric Gustafsson
Monica Rosén (Eds.)

Educational Policy Evaluation through International Comparative Assessments

Waxmann 2014
Münster • New York

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in
the Deutsche Nationalbibliografie; detailed bibliographic data
are available in the Internet at http://dnb.d-nb.de

Print-ISBN 978-3-8309-3091-4
E-Book-ISBN 978-3-8309-8091-9

www.waxmann.com
info@waxmann.com

Cover design: Anne Breitenbach, Tübingen
Typesetting: Stoddart Satz- und Layoutservice, Münster
Print: Hubert & Co., Göttingen

Printed on age-resistant paper,
acid-free as per ISO 9706

Contents

PART A: CONCEPTUAL AND METHODOLOGICAL FOUNDATIONS

PART B: EMPIRICAL STUDIES

PART A
CONCEPTUAL AND METHODOLOGICAL FOUNDATIONS

Rolf Strietholt, Jan-Eric Gustafsson, Monica Rosén and Wilfried Bos

Outcomes and Causal Inference in International Comparative Assessments

Abstract

The main aim of this essay is to discuss how international large-scale assessments can be utilized for policy evaluation studies. We overview key findings from previous studies and propose the curriculum as an organizing concept in considering firstly, how educational opportunities are provided to students around the world, and secondly, the factors that influence how students use these opportunities. Thereafter, we discuss recent developments in the design of the international studies and methodological advances that allow for robust inferences about the causal mechanisms that cause the observed differences in student outcomes. Finally, we identify major challenges for future research including the demand for studies that utilize the trend design of modern studies, more focus on educational equity, and strengthening interdisciplinary and intersectoral collaborations.

Introduction

One of the most salient findings from the field of education is that there are huge national differences in student achievement observed in international comparative studies (Gustafsson & Rosén, this volume). The shockingly large gap between the highest performing countries (mostly in East Asia) and many European countries corresponds to a difference in attainment of two years of schooling. Although this finding has been replicated in several studies (Mullis, Martin, Foy, & Arora, 2012; Mullis, Martin, Foy, & Drucker, 2012; Mullis, Martin, Foy, & Stanco, 2012; OECD, 2014), reasons for and consequences of such differences are currently not well understood. To understand the great need for research in this area it is worth recapitulating some major empirical results stemming from the international comparisons.

As has already been noted, one of the most striking results is the very large difference in mean levels of educational achievement between countries. In the area of mathematics, for example, the TIMS studies show enormous differences in the mean levels of performance between the highest performing countries (most are East Asian) and the lowest performing ones. In the most extreme cases these differences reach more than three standard deviations (SD); one SD corresponding to the effect of approximately two years of schooling. Even within the group of developed countries, the mean differences between the Asian countries and European countries like Sweden is more than one SD. The PISA studies present a similar pattern of differences in educational achievement, with East Asian countries displaying a large advantage in mathematics and science. These studies also show that the school systems of some Western countries, such as Finland, yield a high level of achievement.

Another important finding is that even though there is a general pattern of stability over time, there are also considerable changes in levels of achievement, which are sometimes dramatic. One example is the sharp decline in levels of achievement in mathematics and science in Norway and Sweden after 1995, which amounts to the effect of one year of schooling. A further example is the rapid increase in the level of achievement within the Finnish system from the 1980s, when achievement was at about the same level as the other Nordic countries, to the extremely high level that the country boasts today.

A third cluster of results is connected to the large differences in educational inequality across countries. These differences can be observed in terms of the dispersion of student test scores and inequality of opportunity by gender, social background and ethnicity. Interestingly, the measures of inequality also differ between domains and they change between primary and secondary schools in the respective countries.

The Curriculum Model in the Multilevel Educational System

Many features of the educational systems affect how students learn. The curriculum, broadly defined, is an organizing concept in considering firstly, how educational opportunities are provided to students around the world, and secondly, the factors that influence how students use these opportunities (Robitaille et al., 1993). The curriculum model has three aspects: (1) the intended curriculum is what national educational policies intend students to learn and how the education system should be organized to facilitate this learning; (2) the implemented curriculum includes how the respective educational organization (e.g. schools) implement such goals, what is actually taught in classrooms and who teaches it, and how it is taught; (3) lastly, the attained curriculum describes what students have actually learned, and what they think about it, as well as the emergence of educational inequality (see figure 1).

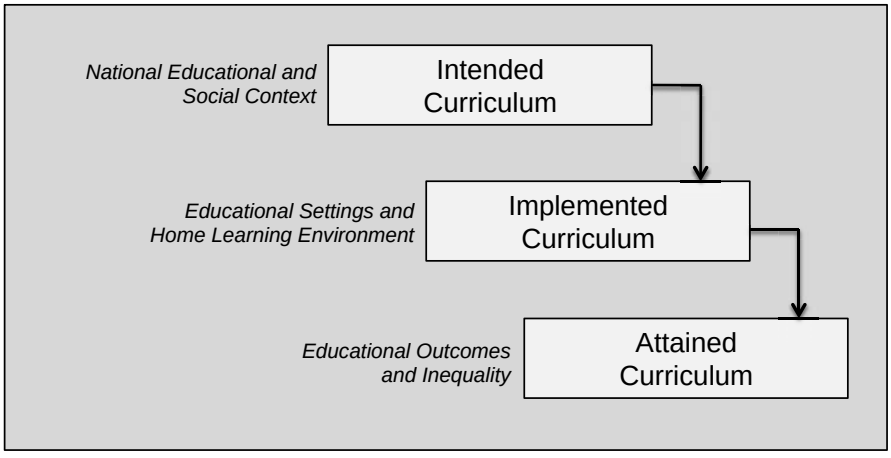

Figure 1: The curriculum model

International large-scale assessments provide a suitable data basis to test the curriculum model because it is possible to uses the achievement tests from international large-scale assessments on student achievement to describe student learning in the participating countries. To form a more complete picture of these learners, information from the international studies' questionnaires (students, parents, teachers, principals) and other relevant sources (e.g. UNESCO Institute for Statistics, OECD statistics) providing a wealth of information.

It is worth recapitulating that educational systems have a multilevel structure where students are nested within classes, classes nested within schools, and schools being nested within regions, societies and nations. Although educational policies are typically located at higher levels they also manifest on lower levels. For this reason it is important to study direct, mediating and moderating effects at the various levels to understand the complex mechanisms within the educational system.

Educational Policy Evaluation through International Comparative Assessments?

Educational effectiveness research aims to understand how and under what circumstances students learn (Creemers & Kyriakides, 2008; Kyriakides & Charalambous, this volume). International comparisons are particularly useful to evaluate the impact of educational reforms and measures. As variation in many system-level features can only be observed across countries (e.g. the existence of central exams), international comparative studies provide a unique approach to study how educational policies and societal issues affect learning and the emergence of educational inequality (Hanushek & Wößmann, 2011). International assessments facilitate comparable measurements of central outcomes of educational systems not only within but also between countries. Since the start of the new millennium, Gustafsson (2008) observes the implementation of a new generation of international comparative studies with a trend design. Recent assessments such as PISA and TIMSS are repeated every few years and thus have a longitudinal component at system level. Unlike previous cross-sectional comparisons, such longitudinal designs allow researchers to estimate causal effects of changes in educational policies and other factors at the system level.

Overview of Development of International Studies. The International Association for the Evaluation of Educational Achievement (IEA) was founded in 1958, with the aim of understanding the factors influencing student achievement. The researchers used the metaphor of using the world as an 'educational laboratory' to investigate the effects of school, home, student and societal factors on educational outcomes arguing that an international comparative approach was necessary to investigate the effects of many of these factors.

During the 1960s and 1970s two main studies were conducted, one on mathematics (First International Mathematics Study (Husén, 1967, Postlethwaite, 1967)), and

one on six different subjects (Six Subject Study (Walker, 1976)). During the 1980s the studies in mathematics, science and reading literacy were repeated (Second International Mathematics Study (Pelgrum, Eggen, & Plomp, 1986), Second International Science Study (Postlethwaite & Wiley, 1992), Reading Literacy Study (Elley, 1992)). While many interesting results were obtained, it was obvious that the studies were not particularly successful at answering questions regarding the determinants of educational achievement, and the causal mechanisms involved. The primary reason for this was that the studies conducted were cross-sectional surveys, and such designs do not easily support causal inference.

In 1995 the TIMS study (Third Mathematical and Science Study), which was a study of enormous scope and complexity, was launched (Martin et al., 1997; Mullis et al., 1997). This study was heralded a major success, and it marked the beginning of a new phase in the development of international studies. In this phase, the presence of educational researchers is less marked and the involvement of national administrative and policy institutions is stronger. Even though researchers are still involved in the design, analysis and reporting of the international studies, the level of ambition in the reporting of important international findings is rather limited. The task of analyzing the factors behind the outcomes for the different countries is left to each participating country, and the databases are made available to the research community for secondary analysis. There has thus been an unfortunate drift away from explanations of causality to the more descriptive aims, mainly serving the purpose of evaluation of educational quality.

Since 1995, the TIMS study has been repeated on a four-yearly cycle, the acronym TIMSS now standing for Trends in International Mathematics and Science Study, and the number of participating countries has increased successively. In 2001, a study on a five-year cycle assessing reading literacy in Grade 4 (PIRLS, Progress in International Reading Literacy Study; Mullis, Martin, Gonzalez, & Foy, 2003) was also established, based upon the same solid design principles as TIMSS.

In 2000, the OECD launched its popular Programme for International Student Assessment (PISA), which covers mathematics, science and reading attainment in 15-year olds (OECD, 2001). PISA includes all the OECD countries, along with a large number of associate countries, and it is repeated every third year. This study uses methods and techniques that are similar to those used in the IEA studies. However, while the IEA studies focus on curriculum defined knowledge and skills, the OECD studies also try to capture competencies expected to be important in adult life. Furthermore, while the IEA studies have a base in communities of researchers, the OECD studies have a more explicit policy-orientation, aiming to influence the educational systems of the member states.

One area that is not well represented in the studies conducted by the IEA and OECD during the last few decades is that of foreign languages. In 2011, however, the European Survey on Language Competences (ESLC) was conducted in 16 European countries and educational entities, and the study, which investigates reading, listening and writing in several languages, has recently been completed (European Commission, 2012).

Research Methodology. During the last two decades, there have been important methodological developments which have made it possible to address issues which were previously impossible to approach. A brief overview of important developments in the fields of measurement and causal inference from observational data is given below.

The fields of educational and psychological measurement have seen remarkable developments in powerful statistical methods through the evolvement of modern test theory or item response theory (Boeck & Wilson, 2004). The power of IRT comes from the fact that parameters of probabilistic models of performance on test items are invariant over samples of persons and items, while the statistics computed within the framework of classical test theory are dependent upon the sample of persons and on which particular combinations of items are used. Since the early 1990s, IRT has been used regularly in the international studies, and through employment of these techniques, the quality of the studies has improved immensely. With the IRT methodology, matrix-sampling models in which different persons take different subsets of items have been implemented, as have methods for equating the scales of different studies.

Another significant contribution to the field of measurement is the development of structural equation and latent variable models (SEM) (Muthén, 2002). Through formulating models in terms of both latent and manifest variables, SEM can deal with errors of measurement in observed variables. Such models can also estimate both direct and indirect effects of chains of variables. Over the last twenty years, SEM has been developed in several different ways, such as for analyzing categorical data, and for addressing the nested structure of units of the educational systems, students being clustered in classrooms, classrooms being clustered in schools and schools being clustered in municipalities, and so on (see Stancel-Piątak & Deana Desa, this volume, for an application). Currently SEM can be employed to model up to three levels of latent variables.

Another important strand of development concerns analytical approaches that allow valid causal inferences based on observational data (Morgan & Winship, 2007, Schlotter, Schwerdt, & Woessmann, this volume). The randomized experiment is a prototypical way to achieve valid conclusions about causal effects, but the challenge is greater when the researcher cannot manipulate conditions in experimental designs. Indeed, many interesting research issues within the field of education are not suitable for experimentation, for ethical, practical and economic reasons. This forces the researchers to rely on different types of observational data. However, a problem with using such data is that associations are not easily interpretable in causal terms, which is to say that with such data it often not possible to say that one factor actually causes a particular outcome. One reason for this is that a variable that is assumed to be dependent may partially cause an effect in a variable that is assumed to be an independent variable. This is what is known as the problem of reverse causality or endogeneity. Another reason why an observed association between two variables need not express a causal relation is that there may be one or more variables that have been omitted from the study, which affect both variables. A further threat that

can be problematic when interpreting results in terms of causality is the threat of errors of measurement in observed variables. Such errors tend to cause systematic underestimation of relations between variables. Several approaches have been developed to guard against the different threats to valid causal inference in analyses of observational data:

- One class of approaches relies on conditioning techniques. The basic strategy is to find a set of control variables that can be included in regression equations in order to remove the effects of omitted variables. The multilevel and SEM approaches allow more efficient and correct analysis of multilevel and error-laden data, and propensity score matching techniques add additional power. However, even though conditioning works well when we have a valid and reliable measure of the control variables, many omitted variables can only be partially observed, and there may be unobserved omitted variables. As such, conditioning is not an infallible approach to arrive at valid causal inference.
- Another approach is instrumental variables (IV) regression. The idea is to find a variable (an 'instrument') that is related to an independent, endogenous, variable X, but not to the dependent variable Y, except indirectly via X (Angrist & Krueger, 2001). The treatment effect is identified through the part of the variation in X that is triggered by the instrument. This approach is often used to deal with problems of reverse causality and errors of measurement and there are many examples of successful applications, particularly within the field of economics. However, IV regression suffers from limitations as well. For example, the standard errors of IV estimates tend to be large, and it is based on quite strong and generally untestable assumptions.
- Within social sciences, longitudinal designs are frequently used (Gustafsson, 2010). When the units under study have characteristics that remain constant over time, the units can be used as their own controls, which brings the advantage that fixed characteristics can be omitted without causing any bias. Regression analysis with change scores for independent and dependent variables, or regression with 'fixed effects' in which each observed unit is identified with a dummy variable can be used to conduct such analyses. This approach does not require longitudinal observations at the individual level, but can be applied at other levels of observation.
- Repeated cross-sectional designs that are used in the international studies of educational achievement to measure achievement trends have a longitudinal design at the country level (Liu, Bellens, Van Den Noordgate, Gielen, & Van Damme, and Rosén & Gustafsson, both in this volume, provide examples). Therefore, with data aggregated to the country level it is possible to take advantage of the strength of longitudinal designs. Analysis of longitudinal data at aggregated levels is often referred to as differences-in-differences analysis. Aggregated data also has the advantage that mechanisms, which at the individual level cause reverse causality, need not be present at higher levels of observation. Furthermore, such data is not influenced by errors of measurement to the same extent as individual data. The

downward biasing effect of errors of measurement is much less of a problem using this approach than with individual data.

One of the main criticisms of the international studies is that the varying characteristics of nations in terms of culture, history and populations make it impossible to draw any inferences concerning the causal effects of different aspects of the educational system (Wiseman & Baker, 2005). This criticism basically expresses the problems caused by omitted variables in between-country comparisons, which is correct. Most of these problems are avoided, however, with a country-level longitudinal approach as it is possible to investigate change and development using such a technique.

This description of advances in methodology for making causal inferences from observational data suggests that there are indeed tools available that can be fruitfully applied to investigate substantive research problems within the field of education. It also is clear, however, that used alone each of the different methods have their limitations, which makes it necessary to use multiple approaches, to attend to possible sources of bias, and to find innovative ways to analyze the complex data from international comparative studies.

Conclusions and Challenges

For decades international comparative studies had cross-sectional designs and the possibilities to use such data for studies that aim at identifying the causal effects of educational policies on educational outcomes were limited. It is only within the last 10–15 years that studies with a longitudinal trend component have been implemented. New data from multiple cycles of such studies are now available. It seems to be a promising approach that future research makes use of the fact that such newly available data from trend studies are much more appropriate to use when testing hypotheses about the causal effects of certain educational policies and reforms on student learning than cross-sectional data.

Most previous educational effectiveness research focused on average levels of achievement. A challenge for future research is to go beyond the currently dominant focus on averages in educational outcomes by emphasizing the idea that educational equality is an equally important outcome of educational systems (Strietholt, this volume). Different ways to operationalize equality and inequality have to be considered and discussed in terms of underlying theories of justice.

Furthermore, the consolidation of various disciplines promises to generate new multidisciplinary approaches to educational effectiveness. Traditionally, economists, sociologists and political scientists typically investigate social structures, institutions and other phenomena that are located on higher levels of the educational system. Conversely, educational scientists and psychologists typically are concerned with individual differences, and therefore focus their attentions on the lower levels of the system, namely the individual students, educators or principles. The dif-

ferent research traditions are also visible in the various methodological approaches that have traditionally been used. On the one hand, econometrics has a particular strength in estimating causal effects from observational data. On the other hand, psychometricians and educational measurement experts have developed elaborated models to test competences and attitudes. From the point of view of research on educational policies and their effects on student learning, it is, however, necessary to attend to both individuals and institutions, and to take account of the multi-level nature of educational phenomena.

Finally, we feel that it is worth to strengthen collaborations between public and private organizations. It is quite obvious that the integration of different sectors is less developed in education in comparisons to other scientific field like engineering or pharmacies. In this context, it is important to mention that it is not universities but organizations like the ACER (Australian Council for Educational Research), ETS (Educational Testing Service), IEA (International Association for the Evaluation of Educational Achievement), and the OECD (Organisation for Economic Co-operation and Development) that are internationally responsible for almost all large-scale studies on student achievement that have been carried out to date. They are the driving forces behind the development of new survey and testing methodologies and for the implementation of new studies. However, these organizations tend to produce reports that merely describe international differences in educational achievement without explaining what the root causes of such differences are. The collaboration of the private sector with leading university researchers might strengthen future international studies as research institutes can engage with private sector partners not only in describing international differences but also in explaining the causes of them. At the same time the collaboration promised to enhance the capacities of universities to conduct international comparative studies. University researchers or groups of researchers from different universities may, for instance, make use of the existing infrastructure of studies like PISA and TIMSS by adding national extensions (e.g. adding an individual panel). This would be a valuable resource for those researchers wishing to answer specific questions particular to their nation's educational systems.

References

Angrist, J. D. & Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives, 15*(4), 69–85.

Boeck, D. & Wilson, M. (2004). *Explanatory item response models: a generalized linear and nonlinar approach*. New York: Springer.

Creemers, B. P. M. & Kyriakides, L. (2008). *The dynamics of educational effeciveness*. London: Rouledge.

Elley, W. B. (1992). *How in the world do students read? IEA Study of Reading Literacy*. The Hague: IEA.

European Commission. (2012). *First european survey on language competences: Final report*. Luxembourg: Publications Office of the European Union.

Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal, 7*(1), 1–17.

Gustafsson, J.-E. (2010). Longitudinal designs. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological Advances in Educational Effectiveness Research* (pp. 77–101). London and New York: Routledge.

Hanushek, E. A. & Wößmann, L. (2011). The economics of international differences in educational achievement. In E. A. Hanushek, S. Machin & L. Wößmann (Eds.), *Handbook of the economics of education.* (Vol. 3). Amsterdam: Elsevier.

Husén, T. (Ed.). (1967). *International study of achievement in mathematics: A comparison of twelve countries* (Vols. 1–2). Stockholm: Almqvist & Wiksell.

Martin, M. O., Mullis, I. V. S., Beaton, A. E., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1997). *Science Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study (TIMSS)* (Vol. Chestnut Hill, MA): Boston College.

Morgan, S. L. & Winship, C. (2007). *Counterfactuals and causal inference. Methods and principles for social research*. Cambridge: University Press.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1997). *Mathematics Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study (TIMSS)* (Vol. Chestnut Hill, MA): Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 International Results in Science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Foy, P. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools in 35 countries*. Chestnut Hill, MA: Boston College.

Muthen, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*(1), 81–117.

OECD. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris: OECD.

OECD. (2014). *PISA 2012 Results: What Students Know and can do Student Performance in mathematics, reading, and science (volume 1, revised edition, February 2014)*: OECD Publishing.

Pelgrum, W. J., Eggen, T., & Plomp, T. (1986). *Second International Mathematics Study: The implemented and attained mathematics curriculum – a comparison of eighteen countries*. Washington, DC: Center for Education Statistics.

Postlethwaite, N. (1967). *School organization and student achievement: A study based on achievement in mathematics in twelve countries*. Stockholm: Almqvist & Wiksell.

Postlethwaite, T. N. & Wiley, D. E. (Eds.). (1992). *The IEA Study of Science II: Science achievement in twenty-three countries*. Oxford: Pergamon Press.

Robitaille, D. F., Schmidt, W. H., Raizen, S., McKnight, C., Britton, E., & Nicol, C. (1993). *Curriculum frameworks for mathematics and science: TIMSS monograph no. 1*. Vancouver, Canada: Pacific Educational Press.

Walker, D. A. (1976). *The IEA Six Subject Survey: An empirical study of education in twenty-one countries*. Stockholm: Almqvist & Wiksell.

Wiseman, A. W. & Baker, D. P. (2005). The worldwide explosion of internationalized educational policy. In D. P. Baker & A. W. Wiseman (Eds.), *Global trends in educational policy* (pp. 1–26). Oxford: Elsevier.

Jan-Eric Gustafsson and Monica Rosén

Quality and Credibility of International Studies

Abstract

Large-scale survey studies of educational achievement are becoming increasingly frequent, and they are visibly present in both educational policy debates and within the educational research community. These studies face a large number of methodological challenges which, in combination with the fact they often yield unpopular results, are reasons why these studies are frequently contested on quality grounds. Taking starting points in two published papers criticizing international studies, methodological challenges related to the validity of the paper and pencil based measurement instruments and to the applicability of the scaling models based on item response theory, are discussed. It is concluded that, while international studies do indeed face methodological challenges that need further work, there is little reason to reject the studies as yielding invalid results on the basis of the expressed criticism.

International comparative studies of educational achievement currently form one of the most conspicuous phenomena in the field of education. At an increasing rate, these studies produce data that policy-makers can worry about and take advantage of, and that researchers can use in analyses of achievement differences between and within countries, and as a basis for investigating effects of different educational and societal factors on educational achievement. Such international studies are, furthermore, a hotly debated phenomenon (e.g., Hopmann et al., 2007; Novóa & Yariv-Mashal, 2003; Simola, 2005), which attracts considerable media attention and which has profound influence on educational policies.

One fundamental question that is often raised in relation to international studies is whether the results that they present can be trusted. Some researchers are likely to respond to this question with a very definite 'no' while other researchers are likely to respond to this question with an equally definite 'yes'. However, international comparative studies are extremely complex endeavors, so it would seem unlikely that their results can be blindly trusted and, given the large amount of resources spent on them, it would also seem unlikely that they are completely untrustworthy. The purpose of the present chapter is, therefore, to discuss some recent methodologically-oriented challenges to the quality and credibility of international studies.

One line of criticism concerns the measurement design of international studies and argues, basically, that with paper and pencil tasks it is not possible to obtain valid results concerning students' knowledge (e.g., Schoultz, Säljö and Wyndhamn 2001). The other line of criticism is raised from a quantitative methodological point of view, questioning the ways in which measurement and scaling models based on item response theory are being applied (e.g., Kreiner & Christensen, in press). Both

lines of criticism are in a sense devastating because, if the critiques are correct, international studies are fraught with fundamental problems which invalidate the entire approach. These two lines of criticism will be focused upon below. First, however, there is reason to provide some background on the development of international studies.

Background and Development of International Studies

The International Association for the Evaluation of Educational Achievement (IEA) was founded in 1958 by a small group of educational and social science researchers, with the purpose of conducting international comparative research studies focused on educational achievement and its determinants. Their aim was to understand the great complexity of factors influencing student achievement in different subject matter domains (Husén & Postlethwaite, 1996; Papanastasiou, Plomp, & Papanastasiou, 2011). They used the metaphor that they wanted to use the world as an educational laboratory to investigate effects of school, home, student and societal factors, arguing that an international comparative approach was necessary to investigate effects of many of these factors. The researchers also had the responsibility of raising funding and conducting the entire research process, from theoretical conceptions and design, to analysis and reporting.

The TIMSS (Third International Mathematics and Science Study) 1995 study (Beaton et al., 1996) marks the beginning of a second phase in the development of international studies (Gustafsson, 2008). Now, the researcher presence is less marked and there has been a shift away from explanatory towards descriptive purposes. The involvement of national administrative and policy institutions has become stronger and, even though researchers are still involved in the design, analysis and reporting of the studies, the level of ambition of the reporting typically is limited. International reports mainly describe outcomes, along with background and process factors, but there is no attempt to explain the variation in outcomes between school systems, or to make inferences about causes and effects. The task of analyzing the factors behind the outcome for different countries is left to each participating country, and the databases are made available to the research community for secondary analysis. Thus, there has been a drift from explanation to description, mainly serving the purpose of evaluation of educational quality as a basis for national discussions about educational policy.

After 1995, there also has been a dramatic increase in the volume and frequency of studies. The number of countries participating in a particular study has increased dramatically and now often amounts to more than 60 countries or school systems. The frequency of repetition has also increased: the IEA studies of mathematics and science (i.e., TIMSS), and reading (i.e., PIRLS), are now designed to capture within-country achievement trends and are therefore repeated every fourth or fifth year. The OECD PISA study ("Programme for International Student Assessment", which cov-

ers mathematics, science and reading, includes all the OECD countries, along with a large number of associate countries, and is repeated every third year.

There were several reasons for this upsurge of interest in international comparative studies in the 1990s. One was that, since the 1980s, there has been an increased focus on outcomes of education, partly as a consequence of the changes in educational governance through processes of decentralization and deregulation. Another reason was that great advances had been made in the methodology for large-scale assessment of knowledge and skills. International studies adopted the methodology developed in the National Assessment of Educational Progress (NAEP) in the United States in the 1980s, based on complex item-response theory, matrix-sampling designs and sophisticated stratified cluster sampling techniques (Jones & Olkin, 2004). This methodology was well suited for efficient and unbiased estimation of system-level performance, and it was skillfully implemented to support international studies. The TIMSS 1995 study was the first study to take full advantage of this technology and, when PISA started a few years later, similar techniques were adopted in that study.

Stability of Results in International Studies

It does seem reasonable to assume that, unless the technology implemented in the TIMSS 1995 study and later studies had generated results that were perceived as being trustworthy, the great boom of international studies would not have taken place. By and large, it seems that country-level results keep quite stable over time. Even though this is of course not necessarily a demonstration of reliability, a pattern of random variation in the outcomes for different countries over time would cause stakeholders to lose faith in the studies.

There also are several examples of countries that repeatedly perform unexpectedly poorly or unexpectedly well, and where it rather seems that expectations were incorrect, compared to the measured outcomes. One example of unexpectedly high levels of achievement is provided by the excellent results of East Asian countries, surprising given that the Western literature had indicated that instructional practices in East Asia were traditional and backward, failing to keep pace with the latest development in learning and instructional theories (Leung, 2008). Another, similar, example is Finland where the PISA results have been unexpectedly high.

The stability of both expected and unexpected outcomes suggests that there must be at least a basic level of quality and credibility in the international studies, as does the expansion of the studies. However, this has been bought at the price of having adopted a very complex technology, inaccessible to educational researchers and policy-makers, and which even very few specialists master. Furthermore, the international comparative studies on student achievement have a somewhat deceptive appearance. They involve students who work on tasks that are similar to those used in classrooms in everyday schoolwork. Yet, the primary purpose is not to provide knowledge about everyday classroom activities but to make generalized descriptions of achievement outcomes at the school system level.

These studies also have the appearance of research studies, involving large and representative samples of students, teachers and schools, and a large number of instruments designed to capture not only student outcomes but also many categories of background and explanatory variables. Yet, they are not designed to test theories or provide explanations, but rather to provide an infrastructure for research through generating data that may be used to investigate a wide range of issues.

Limitations of Paper and Pencil Assessments

Typically, items in international studies are presented in written form and require written responses. Such items are seen by many as artificial and restricted, and it has been argued that more authentic performance assessments should be preferred.

The TIMSS 1995 study offered countries the opportunity to administer a set of performance assessment tasks in science and mathematics to additional samples of students not participating in the main study (Harmon, Smith, Martin, Kelly, Beaton, Mullis, Gonzalez, & Orpwood, 1997). In the study, about a dozen different tasks were administered to students in Grades 4 and 8, each student being given three or four tasks. Altogether, some 20 countries participated in the performance assessment study, even though participation rates were not acceptable in all countries. The overall level of achievement on the performance tasks agreed quite well with the results in the written assessments, though limitations in the data prohibited deeper analyses. Another finding was that countries that did well overall generally tended to do better than other countries on each of the tasks, even though there was also some variation in rank ordering across tasks.

After this first study of performance assessments in international comparisons, no other TIMS study has included such tasks. The reason for this is that they are time-consuming to administer and score whilst, at the same time, the increase in information yield is marginal compared to paper and pencil tasks.

Level of Performance in Paper and Pencil Tests vs. Interviews

However, Schoultz, Säljö and Wyndhamn (2001) argued that paper and pencil tasks have severe limitations, influencing their reliability and validity. They took a starting-point in a socio-cultural perspective and argued that differences in performance should not be seen as a consequence of students' abilities and knowledge; performance should rather be seen as produced through concrete communicative practice. In particular, they argued that there are difficulties associated with the particular communicative format of test items which are presented in written form and which require a written response. They thus claimed that reading and responding to test items in solitude cannot be taken as an unbiased indicator of what students know and understand.

Schoultz et al. (2001) selected two items from the TIMSS 1995 study for scrutiny in an interview study comprising 25 Swedish Grade 7 students. One was an optics item. It presented an illustration showing two flashlights, one with and one without a reflector, and the question was which of the two flashlights shines more light on a wall 5 meters away. An open response was requiredand, to be scored correct, the response had to include an explanation that argued that the reflector focused the light on the wall.

According to the TIMSS results, this item was quite difficult. In the Swedish Grade 7 sample, only 39 % of the students answered the item correctly, which figure was somewhat below the international average. In the interview study, 66 % of the students gave correct answers. Even though this small and possibly unrepresentative sample makes it difficult to compare this result with that from the TIMS study, it nevertheless indicates that the interview situation makes the item easier. One reason for this was that the students did not have to write the answer in the interview situation. Furthermore, many students did not understand the word "reflector" and so had initial difficulties connecting what was written in the question with the illustration but, in the dialogue with the interviewer, these things were clarified. Thus, the higher performance in the interview study was to a large extent due to the scaffolding provided by the interviewer in a Socratic dialogue.

The other item was a multiple-choice chemistry item where the results were even more dramatic. According to the TIMSS data, only 26 % of the Swedish Grade 7 students chose the correct response alternative butin the interview study, no less than 80 % of the students responded correctly. Like the previous case, this was due to the interaction between the interviewer and the interviewee thus helping the students to interpret the text and the meaning of the response alternatives.

From this study, the authors concluded, among other things, that the low performance demonstrated in the TIMS study was due to the fact that the students were limited to operating on their own, and in a world of paper. They concluded that: "Knowing is in context and relative to circumstance. This would seem an important premise to keep in mind when discussing the outcomes of psychometric exercises." (p. 234).

This may seem to be a serious criticism, not only of the TIMS study, but also of results from paper and pencil tests generally. However, the results of this study have little to do with quality aspects of the TIMSS assessment, or of the validity of paper and pencil tests. The Schoultz et al. (2001) study appears at a surface level to deal with the validity of items in the TIMSS test, but this study has in fact different aims and it is based on different assumptions than those made in TIMSS. As will be shown, this makes it impossible to make any inference about the phenomena studied in TIMSS from the results obtained in the Schoultz et al. study, and vice versa.

The most fundamental difference concerns the assumptions made about the nature of performance differences over different contexts. Schoultz et al. view the performance differences between the paper and pencil and interview situations as absolute while, in TIMSS, performance differences between two situations are seen as relative. They interpret the higher level of performance when the item is admin-

istered in an interview situation compared to a paper and pencil situation as evidence of a higher level of knowledge and conceptual insight, and therefore as better evidence of what students can actually accomplish. This interpretation also implies that, if TIMSS were to use interviews to a larger extent than is currently done, this would result in a more positive picture of student knowledge. However, this is not so because in TIMSS the observed performance level is seen as being determined not only by student ability but also by the difficulty of the item. Thus, a TIMSS researcher who is presented with the finding that the level of performance is higher when an item is presented in a highly supportive interview context than in a paper and pencil context, would not necessarily think that the level of ability becomes higher when students are interviewed than when they sit alone and read and write. Another, more reasonable, interpretation is that the level of ability of the person is more or less constant in the two situations, while the task presented in the interview situation is easier than the paper and pencil task.

Another difference between the assumptions underlying the Schoultz et al. (2001) study and the TIMS study concerns the notions of reliability and validity. Schoultz et al. (2001) argue that it is possible to subject the TIMSS items, which already had been tested for validity and reliability, to a further test which, in a truer sense, would reveal the actual validity and reliability of the items. According to this view, the items have immanent and absolute characteristicswhich can be revealed through a careful and detailed analysis of the context in which the student interacts with the item. This view is related to the absolute view of student performance discussed above. A person working on large-scale assessment would, by contrast, find such a view to be incomprehensible because, according to the assessment view, the constructs of validity and reliability do not primarily refer to characteristics of single items, but to collections of items. Thus, the most commonly used form of reliability refers to the internal consistency of a scale based on many items. Similarly, the most fundamental concept of validity, namely construct validity (Messick, 1989), is not applicable to an item in isolation.

When it comes to reliability and validity, it would rather seem that the Schoultz et al. (2001) study faces serious problems making credible inferences about students' absolute level of ability to perform the two tasks on the basis of an interview study which, in many respects, was more like a teaching situation than a testing situation.

According to this analysis, Schoultz et al. (2001) have made the mistake of starting from one set of assumptions, which emphasize the context-bound nature of human action and interaction, and have applied them to an activity which is based on the assumption that it is possible to generalize across contexts to the system level. This generates more confusion than clarification because concepts and observations that seem to refer to the same phenomena do, in fact, refer to different phenomena.

The problem is that Schoultz et al. (2001) have applied the socio-cultural perspective with its set of assumption in a critique of a phenomenon that is based on quite different assumptions. Their results therefore do not invalidate the TIMS study; nor is it possible to argue that the present criticism invalidates the socio-cultural approach in general. It might, however, be worthwhile trying to capture the differ-

ence between the two perspectives in somewhat more constructive terms than just to state that they are different. One way to capture different perspectives is to describe them in terms of metaphors. So let us introduce a metaphor intended to do just that.

Weather and Climate

We are almost always concerned with weather because it profoundly affects our daily life; decisions about what clothes we should wear, if we should go to the golf course or to the museum, if it would be advisable to take the car or not, just to mention a few examples. Weather also affects our mood, and it supplies us with conversation material in almost all social contexts. However, we cannot do much about the weather, except adapting to the conditions it creates for us. Fortunately, meteorologists can predict what the weather will be like within the next couple of days. However, there is a margin of error in these predictions and, beyond a week or so, the predictions are useless. This is because of the great complexity of weather phenomena, and because the weather is chaotic; it is not even theoretically possible to predict weather over longer periods of time.

Should we not like the weather there is not much to do, except, of course, to move to a place with a better climate. Simple indicators, like average temperature, average rainfall, and number of days with sunshine, give us much information on which to compare the climates of different places. However, even though such information tells us much about the climate, they do not tell us much about what weather we are likely to experience on a particular visit, because these numbers are averages with a lot of variation. Thus, the link between climate and weather is a weak, probabilistic, one.

However, while weather is unpredictable and chaotic, climate and climate changes are stable phenomena which we can understand theoretically and for which empirically-based models, predicting long-term development, can be constructed. It could be argued that climate does not exist, in the sense that we cannot experience it directly. We do experience weather, however, and through aggregating these experiences, we get a sense of climate. In a more precise manner, scientists define climate as aggregate weather, using indicators such as mean temperature. Thus, climate is an abstraction which, in a sense, only exists in theoretical models. Nevertheless, it is a powerful abstraction which has very concrete and important implications for how we could and should live our lives.

In terms of this metaphor, large-scale survey studies are concerned with climate, while research that focuses on context-bound phenomena is concerned with weather. Thus, the assessment in TIMSS is based on aggregation of a very large number of item responses, little or no interest being focused on the particular items. In contrast, the Schoultz et al. (2001) study is focused on particular contexts.

Many object to aggregation of observations in educational and psychological research, ascribing validity only to that which can be directly observed (e.g., Yanchar & Williams, 2006). But the argument can also be turned around and it can be argued

that, in order to see the general aspects (e. g., the climate), it is necessary to get rid of the specifics (e. g., the weather). Seen from this perspective, methods which conceal context-dependent variation have strengths, rather than disadvantages, when the purpose is to investigate general patterns and relations.

Low- and High-level Inference Research

The Schoultz et al. study would be classified as a qualitative study, while the TIMS study would be classified as a quantitative study. However, Ercikan and Roth (2006) challenged the meaningfulness of this distinction, arguing that the quantitative and qualitative dichotomy is fallacious. One of their arguments was that all phenomena involve both quantitative and qualitative aspects at the same time. As an alternative to the quantitative/qualitative distinction, Ercikan and Roth (2006) proposed that different forms of research should be put on a continuous scale that goes from the lived experience of people on one end (low-level inference) to idealized patterns of human experience on the other (high-level inference). According to Ercikan and Roth (2006), "Knowledge derived through lower-level inference processes … is characterized by contingency, particularity, being affected by the context, and concretization. Knowledge derived through higher-level inferences is characterized by standardization, universality, distance, and abstraction … The more contingent, particular, and concrete knowledge is, the more it involves inexpressible biographical experiences and ways in which human beings are affected by dramas of everyday life. The more standardized, universal, distanced and abstract knowledge is, the more it summarizes situations and relevant situated knowledge in terms of big pictures and general ideas." (p. 20)

This level-of-inference approach to characterizing different forms of research is much more useful than the qualitative/quantitative dichotomy. Thus, while research on weather and climate cannot easily be characterized with the quantitative/qualitative distinction, research on weather may be meaningfully described as low-level inference and research on climate as high-level inference. Similarly, the Schoultz et al. (2001) study is an example of low-level-inference research, while the TIMS study is an example of high-level-inference research.

Quality Aspects of High-level Inference Data

While the low-level inference approach can be grounded in interpretations generated from observations in specific contexts, this is not possible in the high-level inference approach. In this approach, the intention is to capture abstractions which span specific contexts and contents. The question, then, is if this is possible and meaningful, and what criteria we can used to decide whether it is meaningful.

Ocular inspection of the items obviously cannot be used, and the answer cannot be found in detailed analyses of the contents and contexts of specific items, even