Document généré le 29 juin 2024 15:20

Mesure et évaluation en éducation

Validation of competence models for developing education standards
Methodological choices and their consequences

Erich Ramseier

Volume 31, numéro 2, 2008

URI : https://id.erudit.org/iderudit/1025006ar
DOI : https://doi.org/10.7202/1025006ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Ramseier, E. (2008). Validation of competence models for developing education standards: Methodological choices and their consequences. *Mesure et évaluation en éducation*, *31*(2), 35–53. https://doi.org/10.7202/1025006ar

Résumé de l'article

À la suite de PISA 2000, la Suisse développe des standards de formation contraignants pour la scolarité obligatoire dans les trois régions linguistiques (projet HarmoS: Harmonisation de la scolarité obligatoire). Ces standards pour quatre sujets sont basés sur des modèles de compétence. Une partie du développement consiste en une étude empirique visant à valider ces modèles de compétence. Le présent article décrit la conception, les méthodes et certains résultats de cette étude de validation. Il traite aussi de l'utilité de l'étude dans l'optique d'une définition des standards ainsi que des conséquences de certains choix méthodologiques, en particulier la façon heuristique d'appliquer le modèle de Rasch.

Tous droits réservés © ADMEE-Canada - Université Laval, 2008

Ce document est protégé par la loi sur le droit d'auteur. L'utilisation des services d'Érudit (y compris la reproduction) est assujettie à sa politique d'utilisation que vous pouvez consulter en ligne.

https://apropos.erudit.org/fr/usagers/politique-dutilisation/

Cet article est diffusé et préservé par Érudit.

Érudit est un consortium interuniversitaire sans but lucratif composé de l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal. Il a pour mission la promotion et la valorisation de la recherche.

https://www.erudit.org/fr/

Validation of competence models for developing education standards: Methodological choices and their consequences

Erich Ramseier

Cantonal ministry of education, Bern, Switzerland

KEY WORDS: Validity, competence, education standards, Rasch model, dimensionality, differential item functioning

Following PISA 2000, Switzerland launched a project called HarmoS (Harmonization of obligatory School) to develop binding education standards for compulsory education in the three language regions of the country. These standards for four subject areas are based on models of competence. Part of the development is an empirical study to validate these competence models. The present article describes the design, methods, and some results of this validation study. It also discusses the study's usefulness for defining the standards and the consequences of some methodological choices, particularly the heuristic application of the Rasch model.

MOTS CLÉS: Validité, compétence, standards de formation, modèle de Rasch, dimensionnalité, fonctionnement différentiel d'items

À la suite de PISA 2000, la Suisse développe des standards de formation contraignants pour la scolarité obligatoire dans les trois régions linguistiques (projet HarmoS: Harmonisation de la scolarité obligatoire). Ces standards pour quatre sujets sont basés sur des modèles de compétence. Une partie du développement consiste en une étude empirique visant à valider ces modèles de compétence. Le présent article décrit la conception, les méthodes et certains résultats de cette étude de validation. Il traite aussi de l'utilité de l'étude dans l'optique d'une définition des standards ainsi que des conséquences de certains choix méthodologiques, en particulier la façon heuristique d'appliquer le modèle de Rasch.

Author's note – I wish to thank the reviewers for their valuable feedback and Nathan Zoanetti, Assessment Research Centre, University of Melbourne, for his important suggestions. All correspondence should be addressed to Dr. Erich Ramseier, Bildungsplanung und Evaluation, Erziehungsdirektion des Kantons Bern, Sulgeneckstr. 70, 3005 Bern, Switzerland. E-mail: [erich.ramseier@gmx.ch].

Palavras-chave: Validade, competência, referenciais de formação, modelo de Rasch, dimensionalidade, funcionamento diferencial de itens

Na sequência de PISA 2000, a Suíça desenvolveu referenciais normativos de formação para o ensino obrigatório nas três regiões linguísticas (Projecto HarmoS: Harmonização da escolaridade obrigatória). Estes referenciais para quatro domínios são baseados nos modelos de competências. Parte do desenvolvimento consistiu num estudo empírico para validar esses modelos de competências. O presente artigo descreve a concepção, os métodos e certos resultados deste estudo de validação. Trata também da utilidade do estudo na óptica de uma definição de referenciais, bem como as consequências de certas escolhas metodológicas, em particular o modo heurístico de aplicar o modelo de Rasch.

Introduction

PISA, the “Programme for International Student Assessment”, has received enormous public and political attention in Switzerland. After Germany, Switzerland had the second largest press coverage of all countries when the PISA 2000 results were released (Network-A/INES/OECD, 2004). This is probably due to the fact that both countries have a federal structure and did not yet have any system of large-scale student assessment.

After PISA 2000, the Swiss conference of cantonal ministers of education launched a program in 2003 to improve schools, including the definition of education standards and the regular monitoring of the education system. It is an open question to what extent the introduction of monitoring and education standards was a direct political consequence of PISA or whether PISA merely provided timely justification of an already arising political agenda. In any case, Switzerland is now engaged in the project called HarmoS (**Harmo**nization of obligatory **S**chool). Among other objectives, this project develops education standards which are based on models of competence. Part of the developmental process is an empirical study which aims to validate these competence models.

The present article describes the design, methods, and some results of this validation study in the context of the HarmoS project. Moreover, it discusses to what degree the study’s aims are attained and the consequences of some methodological choices.

HarmoS: a political and a scientific project

HarmoS is a multi-part project with the aim to harmonize the 26 cantonal school systems. At the political level, it consists of an intercantonal agreement which defines basic features of the education system including the school enrolment age and the duration of schooling. It also introduces education standards and the monitoring of the attainment of these standards.

Concerning the standards, the intercantonal agreement only states that binding education standards are to be applied. Of course, to implement these standards, they first have to be defined and adopted by the community of cantons. In a first part of this political process, it was decided that education standards must be met by the end of grades 2, 6, and 9. For a start, education standards are defined for the language of instruction, foreign languages (a second national language and English), mathematics, and natural sciences. These education standards have to be measurable and controllable indications of competences which are independent of the curriculum and based on a comprehensive competence model (EDK/CDIP, 2005). Embedded in this political process, a scientific project was initiated to develop and propose such standards by the end of 2007. In the beginning of 2008, most of these proposals entered the administrative, political and public debate which may lead to the official adoption of the standards.

An expertise about the development of national education standards (Klieme et al., 2003) was highly regarded in Germany and Switzerland. Swiss authorities used this expertise as a guideline. The project HarmoS therefore follows a number of principles:

- The concept of competence follows the definition of Weinert (2001). Competencies comprise the mental conditions necessary for solving problems in a socially defined field, e.g. a school subject. They can better be described by the typical demands of that field than by enumerating underlying psychological processes. Competencies are learned and based on content-specific knowledge. They include motivational as well as ethical components.
- Education standards are basic standards that describe what each student should master.
- Standards are embedded in a model of competence. The model of competence has to be substantiated by concrete tasks and tests measuring this competence.

- The empirical validation of the competence model is included in the process of developing the standards. A limited large-scale assessment called a "validation study" is necessary for this. It should be noted that competence models should not be restricted to only those components which can be validated at this stage.

For each subject, a consortium had been designated as responsible for the development of standards. The consortia were composed of experts from the German, the French, and to some extent the Italian language region; many members were didactical experts from teacher training colleges. The consortia were complemented by a methodological group[1] which was responsible for the design of the validation study and gave guidance and assistance regarding validation.

The inclusion of the scientific project into a policy-making process imposed a very tight timeline upon it: consequently, a validated competence model and a proposal of basic standards had to be delivered within eight months after the data collection of the validation study.

In view of the future common use of the four competence models, the consortia agreed on a shared basic structure of the models (*cf.* Figure 1): Two dimensions define the components of competence. One dimension describes content areas, such as shape and space or numbers and variables in mathematics, or general activities like writing or listening for languages. The other dimension describes processes, actions or aspects of competence, such as operations and calculation or reasoning and justifying in mathematics. A third dimension describes the degree of attained competence with levels characterized by typical cognitive processes or performances mastered at each level. Since the competence models have to integrate competencies shown at grades 2, 6, and 9 and to give guidance for teaching these subjects, they should represent the expected competence development across this entire age range. Tasks which can be mastered thanks to the competence in question can be allocated to a specific combination of content area, aspect, and level.

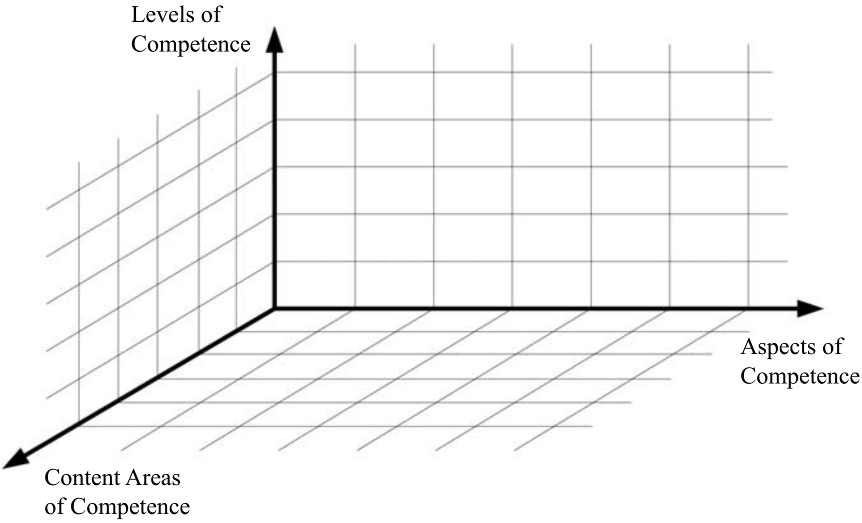

Figure 1. ***Basic structure of HarmoS competence models***

The competence models for the four subjects differ in the way they operationalize this basic structure (*e.g.*, the relation between competence levels and grades). For instance in the mathematics competence model, “can-do” statements for each grade describe the competence at all intersections of a content area and an aspect, whereas levels of the competence within the grades are characterized for each aspect independently of the content areas. Motivational, ethical, and social facets are seen as important attributes of the overall competence, but their representation differs between the consortia and therefore they are not included in the common basic structure of the competence models.

Validation Study: Design

Aim

The aim of the validation study is to examine the emerging competence models empirically. Validation as the quest for validity usually relates to a measurement procedure or to a research design or process. Because competence models have a different status than measurement procedures, it is not obvious what their validation means. A competence model can be seen as the theoretical basis of measurement procedures. Therefore, its validity is linked to the validation of corresponding measurement procedures. This means transforming the competence model into an achievement test composed of concrete tasks and validating this test and its structure. Such validation is a long-term process. The present validation study can only deliver some initial elements, including:

- eliminating tasks with insufficient psychometric quality;
- identifying the empirical difficulty of tasks as an aid for illustrating competence levels through tasks and their typical characteristics at certain levels;
- examining whether the competence model conceived as independent from different cultural, linguistic, and instructional traditions can be maintained (differential item functioning);
- analyzing sub-dimensions corresponding to the proposed components of competence;
- determining the competence distribution in the student population and the percentage attaining the proposed basic standards.

General Design

The present validation focused on grade 6 and 9. The populations were students of these grades enrolled in public schools. Students with special needs who follow a reduced curriculum were included since this group is critical when the study is about basic standards.

The aims of the study require the analysis of a large number of tasks. Many tasks in each subject are needed to illustrate the several content areas and aspects of competence as well as the intended levels of competence. Insufficient pilot testing means that an even larger number of tasks are needed, since many might turn out to be uninformative or poorly related to the construct of competence. From the number of tasks needed, the planned sample size was determined based on the mean time needed to solve a task, the length of testing time per student, the expected response rates of schools and students, the acceptable standard error of item difficulty in each of the three language regions, while taking into account the planned scaling method. Given that students were tested on two days, a planned gross sample size of 6 600 per grade was deemed to be sufficient for first language and mathematics. The sample is by design smaller for science and foreign languages with fewer tasks to be evaluated.

To be able to estimate the competence distribution, the sample had to be representative. A two stage sample was used; in the first stage schools were sampled, in the second two whole classes within a school. Renaud (2006) described the general sample design, Ramseier and Moreau (2007) presented within-school sampling and determination of student weights. The attained overall student response rate was 82% in grade 6 and 85% in grade 9.

Test Design

The four consortia were responsible for developing the tasks and for most aspects of test construction such as content coverage and choice of item formats, assisted therein by a guideline of the methodological group. The tasks were originally developed in German or French and had to be translated into the other language and – for first language and mathematics – into Italian. The consortia assigned each task to one cluster, each cluster taking 20 or 30 minutes of testing time. In total, they produced 96 clusters for grade 6 and 117 for grade 9, all prepared in the two or three languages.

Students were tested on the first day for a total of 80 minutes in mathematics and science and on the second day for a total of 100 minutes in first and second language. Both testing sessions consisted of four consecutive sections. In each test section, students could work on one cluster. This means that each student only worked on eight of the approximately 100 clusters per grade. Therefore more than 90% of all information in the complete data matrix (items by students) is missing by design. It was therefore a major challenge to distribute and combine clusters in such a way that all tasks of a subject could be scaled on a common scale, that correlations between content areas of competence and between aspects of competence could be estimated, and that position effects could be controlled at least between clusters.

The printing arrangement allowed for assigning booklets individually to persons and test sections. On this basis, a balanced design was created by printing each cluster in a separate short booklet and by randomly assigning these booklets to persons and test sections, thereby keeping a few restrictions depending on the needs of each subject. For this assignment, a list including all possible cluster combinations was constructed and matched to the stratified list of students. In total, about 100 000 individual links between clusters, test sections, and persons were produced.

As a consequence of this design, many different individual combinations of testing materials were distributed. For instance, in the German part of Switzerland 2 079 different combinations of mathematics and science clusters entered analysis in grade 9, of which only 10 were dealt with by two students. Given such diversity, possible interactions between clusters are optimally controlled and tasks are linked with each other in many different ways. The design is almost equal to a full balanced incomplete block design which is generally considered as unrealistically complex (Mazzeo, Lazer & Zieky, 2006, p. 685).

Validation study: Scaling and Results

For scaling, item response theory (IRT) was chosen since this psychometric framework suits the study's goals and conditions: IRT explicitly links behavior on tasks to a latent trait, which corresponds to the theoretical concept of competence as a construct and is not directly observable. Specifically, IRT places items with their difficulty and persons with their competence (measured as a latent ability) on the same scale. This allows illustrating and interpreting levels of competence by the characteristics of corresponding tasks (*e.g.*, Embretson & Reise, 2000, p. 25). This is essential since the study mainly aims to describe a competence and not to compare groups of students. For this purpose, it is useful that IRT is a micro theory about scaling and "pays particular attention to items, whereas classical test theory and generalizability theory are largely test-score-based" (Brennan, 2006, p. 6). Finally IRT permits measuring the competence of students on a common scale even though most students have worked on different tasks (Glas, in this issue). Within IRT, the Rasch model was selected because it is clear and simple and allows estimation of item difficulties based on limited sample sizes (Yen & Fitzpatrick, 2006, p. 133). In the simple logistic Rasch model the probability P_{si} of correctly solving a task – instead of failing it – is a function of the difference between the latent competence θ_s of a person s and the difficulty β_i of the task i:

$$P_{si} = \frac{e^{(\theta_s - \beta_i)}}{1 + e^{(\theta_s - \beta_i)}}$$ (*e.g.* Embretson & Reise, 2000, p. 50)

The following results focus on technical aspects and draw on grade 9 mathematics results as an example. The competence model for grade 9 mathematics specifies five content areas and eight aspects of competence; four and six, respectively, could be included in the validation study. More information about mathematical content and examples of tasks is presented in the project documentation (HarmoS Konsortium Mathematik, 2007).

Item selection

Do the items fit the Rasch model and is their discrimination sufficient? To check this, all 269 grade 9 mathematics items were included in a one-dimensional Rasch model estimation using the software ConQuest (Wu, Adams, Wilson & Haldane, 2007). Items were deemed to fit the Rasch model if their infit measure was in the interval 0.7 – 1.3 (*cf.* Wright & Linacre, 1994). Surprisingly, only one item violated this criterion and was excluded; only another two were outside of 0.8 – 1.2. In conclusion, the fit to the Rasch model was quite satisfactory.

Among the remaining items, 42 showed an insufficient classical discrimination with an item-total-correlation below 0.3 and were therefore candidates for exclusion. Since the goal of the study was not to create an efficient test but to illustrate the full range of mathematics competence, 26 of these items were not excluded. These items were typically very easy or very difficult in this population – a condition which leads to a low item-total-correlation even if fit to the Rasch model is satisfactory.

Model adequacy across language regions

The development of competencies depends on learning – a learning process which is shaped by ways of teaching, curricula, the cultural context, and other learning conditions. Due to such factors, task difficulty might vary between subpopulations of students – as accentuated in studies using generalizability theory (Johnson, in this issue): The competence could have different structures in these subpopulations and a common competence model may not be applicable. It is therefore important to analyze differences of item difficulties, particularly between the Swiss language regions with their differences in culture and learning conditions. Another reason for this examination is translation as a possible source of errors.

Figure 2 compares the relative difficulty of the grade 9 mathematics items in the French and German parts of Switzerland. Relative difficulties show the distance between the difficulty of an item and the mean achievement of a region. Despite the apparent unsystematic dispersion in Figure 2, relative item difficulties can generally be considered as similar in the two regions. This unity is shown in a correlation of .91 between the relative difficulties for the two language regions. In a global view, the competence has a similar structure across regions and shows some measurement invariance (Embretson & Reise, 2000, p. 250 f.).

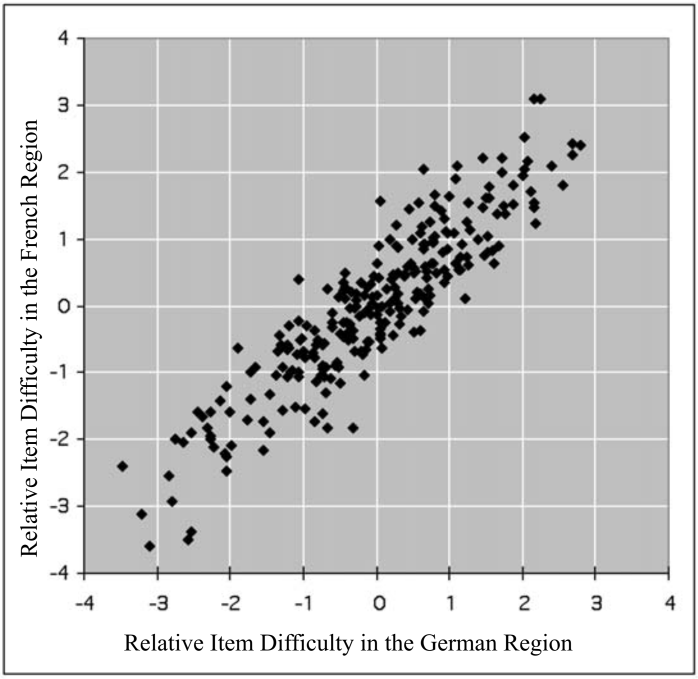

Figure 2. ***Relative Item Difficulty by Language Region***

On the other hand, the relative difficulty of some items with a certain difficulty in one region varies in a range of about 1.5 logits in the other regions. This corresponds to about 1.5 standard deviations in the student population and is considerable. Therefore, items with large differences in relative difficulties (differential item functioning, DIF) were examined before including them in the final scale. Items were identified as critical if they differed in one of the three regions by more than 0.5 logits from the commonly estimated difficulty and if this difference was significant given the standard errors of the difficulty estimates. This fairly lenient criterion (*e.g.*, Tristan, 2006) was applied since this study only begins to validate the competence models and does not judge persons or groups. Fifty-seven of 252 items were identified. Eleven of them were retained for specific content reasons. The other 46 were split into three regional items each: in the following analyses each of these items was treated as three separate items, each only administered in one region. They did therefore not influence comparisons between regions any more but still influenced the competence estimate of persons within a region.

Dimensionality of the competence

Is it possible to empirically distinguish between sub-dimensions which correspond to the four content areas or the six aspects of competence, respectively? Since the competence models assigns each item to a content area, a multidimensional model for content areas could be estimated using these assignments. The same was done for the aspects of competence. The information fit indices BIC and CAIC as well as the chi-square-test show that both multidimensional models describe the data better than the one-dimensional model (*cf.* Table 1). The correlations between the several aspects of competence and between the content areas of competence are quite high. Therefore, it makes sense to regard them as sub-dimensions of an encompassing competence. When interpreting the size of the correlations one has to remember that these coefficients are estimates of correlations between latent variables and not attenuated by measurement error. Hence, they indicate that these constructs are distinguishable. As a point of reference, the correlations between the mathematics subscales in PISA are slightly higher (OECD, 2005, p. 190).

Table 1

Comparison of One-dimensional and Multidimensional Models

						Model comparison[a]		
Model	Correlations	Deviance	df	BIC	CAIC	Diff. of deviance	Diff. of df	p
One-dimensional		132 957	2	132 965	132 967			
Areas of competence	.77 - .83	132 434	14	132 487	132 501	523	12	.0000
Aspects of competence	.78 - .91	132 668	27	132 770	132 797	289	25	.0000

Note: df corresponding to (co)variances and means by fixed item parameters; sample size 5 704; Deviance = -2 log L

[a] Comparison with one-dimensional model; difference of deviance approx. chi-square-distributed (Wu et al., 2007)

Table 1 indicates that correlations between content areas are lower than those between aspects of competence. Therefore, and based on the fit indices, the content area model is preferable if one had to choose between models on an empirical basis. But this is not the case. Both models are interesting views of the mathematics competence. Since they retrieve theoretically formulated dimensions in the data, they both contribute to the validation of the competence model.

Reporting scale, competence levels and basic standards

The scaling conventions used in the model estimation produce person scores varying around zero. In a pedagogical context, such values are unsuitable as a description of competence. To facilitate meaningful interpretations, they have to be transformed. The applied transformation is linear and used for task and person scores, therefore preserving all essential information. As in the PISA study, the convention of transforming Rasch ability measures to a population mean of 500 and a standard deviation of 100 is applied.

The simple logistic Rasch model usually locates items on the scale at the point where persons have a 50% chance of solving them. For pedagogical interpretations, it seems preferable to place items at a point where they are mastered with higher certainty. A chance level of 2/3 was chosen. According to the Rasch model, this implies a common linear shift of all dichotomous items.

To represent higher competence, the conception of the competence model accentuates an ordered sequence of several discrete competence levels rather than a continuous dimension. In the validation study, these ordered competence levels overlay the continuous one-dimensional Rasch scale. The mathematics consortium described four such levels *a priori* and attributed tasks to them based on the expected difficulty of understanding the task, the complexity of cognitive processing, and the complexity of required mathematical concepts and skills. The correlation between the *a priori* attribution to levels and the empirical item difficulty resulting from the validation study was r = .56 in a set of 142 dichotomous items. This degree of agreement is far from being perfect: The empirical difficulty of items is a necessity for properly describing the competence.

For the final competence model, the consortium relied on the ordering of the tasks by their empirical difficulty and revised the description of the levels using typical characteristics of adjacent tasks. This mixture of *a priori* and inductive derivation lead to cutoff points of 400, 540, 635, and 726 on the continuous scale. The varying breadth of the levels is contrary to *a posteriori* defined levels with equal breadth in PISA (OECD, 2005) and illustrates the preference we have given to content rather than formal convenience.

The mathematics consortium had the mission to propose a basic standard. They presented a verbal characterization, illustrated it with selected tasks and assigned the basic standard to the start of the first competence level (cut score 400). This meant that 16 % of the tested population is not reaching this basic standard. The consortium determined the basic standard in an internal consensus-building process without implementing a formal standard-setting procedure like the bookmark method (Zieky & Perie, 2006, p. 20).

Conclusions regarding the HarmoS validation study

The validation study significantly contributes to the development of competence models and basic standards. The empirical anchorage of tasks and their difficulties allows for a substantively meaningful hierarchy of competence levels to be described and illustrated. The theoretical distinction between several content areas and aspects of competence is retrieved in empirical correlations and thereby validated. The possibility and challenge of formulating a common competence model for several language regions is also substantiated. The proportion of students passing the basic standard is known. In sum, the validation study is a necessary part of the development of standards.

On the other hand, the extremely tight timeline associated with this research strongly limited the analyses that could be conducted. Several important questions are not yet answered.

a) The analysis of language differences showed a considerable divergence in the difficulties of items between language regions. A detailed investigation of these differences is necessary to clarify whether they are a consequence of translation problems or whether they indicate substantive differences in cultural traditions or teaching priorities. In the latter case, one would need to discuss consequences : would it be preferable to limit the concept of mathematics competence to a common core, to establish regional competence models, or to adjust regional differences of opportunities to learn in order to foster an encompassing common competence model ?

b) Other forms of differential item functioning (*e.g.*, by school type or gender) have to be analyzed. Yet, it could again be dangerous to eliminate these items since this might change the meaning of the measured concept while items with strong DIF might give hints about what needs to be done in teaching to reduce these differences.

c) The analysis of dimensionality is incomplete. It is probable that more sophisticated models would be superior, for example models which combine content areas and aspects of competence or models which allow assigning items to several competence aspects. Moreover, the optimal allocation of single items is not checked and exploratory analyses could further inform the theoretical development. The multidimensional models show that the condition of local independence (Embretson & Reise, 2000, p. 48) is violated in the one-dimensional model. In this sense, the multidimensional models call into question the practically useful one-dimensional model. Although not fully correct, a parallel use of a general dimension and particular sub-dimensions is not uncommon (Brunner & Süss, 2007).

d) The present validation study focuses on the measurement of competencies at grade 6 and 9, respectively. Therefore, the study examines the appropriateness of the models to represent interindividual differences in the corresponding populations. Because one has to distinguish between interindividual differences and intraindividual changes (*e.g.*, Asendorpf, 1995), the relation between the competence levels and stages of individual development has still to be clarified.

e) The boundaries of the competence levels and the basic standards have been defined by the four consortia without strong support of the methodological group. It might be desirable to consolidate these specifications by applying more explicit, standardized procedures (*e.g.*, Zieky & Perie, 2006).

Altogether, the present study is only an important first step in the process of validating the competence models. The validation must be continued to get a better understanding of the competence models. This process is also necessary to foster the scientific basis for the planned monitoring of the Swiss education system. In a reliable program for examining the attainment of basic standards, pedagogical and didactical as well as methodological aspects and expertise must be integrated. Separate tests in single subjects are not efficient and suitable. In order to establish trends, repeated large-scale assessments must be linked as in PISA. This linking makes high psychometric demands and requires a good fit between the psychometric model and the data. A well-designed assessment system is needed (Scheerens, Glas & Thomas, 2003). This necessitates early and sustained development of the theoretical and methodological basis underpinning the assessment system, including a refinement and extension of the methodological approaches described in this validation study.

Some general considerations about psychometrics in education

The quest for validation of competence models in the HarmoS project raises some issues concerning the role and use of psychometric methods in education that go beyond the context of this specific study. In the following sections, three issues are discussed.

Use of the Rasch model in education

As mentioned earlier, it was necessary to rely on the Rasch model given the goal and conditions of the study. It turned out that very few items had to be eliminated to comply with the assumptions of this model: The choice of the specific scaling model had no strong limiting consequences. This may be a consequence of the preselection of items in the phase of task development (De Pietro, Müller & Wirthner, 2007). Partly, it is due to the liberal application of the selection criteria, especially regarding language differences. A more rigorous approach would be possible (Kubinger & Draxler, 2007). In this project, the application of the Rasch model was not intended as an instance of fundamental measurement (*e.g.*, Andrich, 2004) but rather as a heuristic means for capturing the complex competence in mathematics and the other subjects.

This liberal approach suits the specific measurement conditions in education. For instance in fundamental psychological research, defining and measuring a concept is a purely scientific process within a discipline. There is no harm in revising and differentiating the concept so that it is precisely measurable. In contrast, a concept like mathematics competence in education is socially defined. Even within the scientific discussion, the competence is mainly conceptualized by experts of mathematics education; they give priority to the content of competence and its implication for instruction. The question of measurement only emerges in an interdisciplinary interaction with psychometricians. Furthermore, competence is not a purely scientific concept but is embedded in educational policies and has to prove its utility in the practice of instruction and to adapt to competence requirements of stakeholders in the society and in the economy where these competencies are applied. As a consequence of these social influences, the competence is in constant change.

In this situation, the restrictions imposed by the scaling process have to be as small as possible (Kolen, 2006, p. 156), but strong enough to produce meaningful results. For instance, the Rasch assumption of equal discrimination of items cannot be strictly satisfied by all tasks of a comprehensive competence like mathematics competence. In fact, one has to accept some deviations (Kubinger & Draxler, 2007). Accepting large deviations has its price, though, making the valuable characteristics of the Rasch model only approximately valid. For instance, items are no longer strictly equivalent and a balanced set of items instead of a free selection would be required for valid measurement. The more conclusions there are that depend on a limited number of items and on their model conformity, the more a good fit between model and data is required. Examples of contexts that require good model-data fit include the linking of assessments and adaptive testing systems.

Competence level or competence scale?

Item response theory as well as true score models of generalizability theory and classical test theory describe competence on a continuous scale. In contrast, many theoretical competence models focus on a sequence of distinct levels. In the validation study, levels were assigned to sections on the scale. This was done by an interpretative process based on clustering of items on the scale and their typical characteristics (*cf.* Griffin, 2007). More formal methods could be applied (Rost, 2004). But levels would still only be superimposed on a primarily continuous entity.

Students are attributed to a certain level based on cut scores. For grade 9 mathematics this means that students at the lower end of level 1 (400 points) and those at the upper end are classified as equal although the former only have a chance of 33% to master an item positioned at the upper end of the level, whereas this chance is 67% for the latter. It is questionable that students at the beginning of level 2 with a chance of 68% to master this item are in contrast treated as different. This conceptual limitation perhaps renders such levels more robust for describing the distributed competence of a population rather than specifying the competence of an individual student. Consideration of the consequences of assigning a student to a level adjacent to that which more accurately reflects their true competence is therefore important. So too is consideration of the range of substantive competencies embedded in each level.

Levels of competence are substantial if they represent qualitatively different processes that are needed to master items. For instance, a set of items representing a competence may get much easier once students have acquired a certain concept or procedure. This situation could be represented by characterizing the competence by two classes: concept acquired or not. If the difficulty of other relevant items does not depend on this concept, it would be possible to scale the competence separately for the class of students having acquired the concept and those not having acquired it. The competence would then be described by a mixture of two discrete classes and a continuous ability within a class. Indeed many models exist which allow one to represent a range of such situations (National Research Council, 2001 ; Rost, 2004). However, in the present study the competence in question is highly comprehensive and complex ; many years of instruction and a multitude of acquired concepts and procedures contribute to it. This multitude blurs the influence of a single acquisition and the situation approaches that of a continuous competence and better conforms to the corresponding IRT assumptions.

Therefore, at least for complex school based competences, continuous scales seem most suitable. Superimposed discrete levels are then only a practical means for illustrating the meaning of sectors of this continuum – a means that is nonetheless very important in the communication with the pedagogical community.

Impact of psychometric models in a broader context

The choice of an adequate psychometric model for monitoring the education system is certainly important. The model may shape the competencies measured and have an influence on tests and therefore indirectly on the aims and modalities of education. However, this impact on the daily practice in school and instruction must be put in perspective. The impact of the methodological choice is only a part of the total impact of the monitoring program. The sheer existence, aim and design of this program alone are much more influential : Is it a high stakes program ? Are selection processes or school rankings based on that program ? Is assessment feedback public or confidential ?

Given the aim and function of the program, the impact on educational practice depends on what exactly is measured : Is it about a competence, that is about the ability to apply knowledge to solve new problems in question, or is it just about reproduction of knowledge ? The impact then depends on the trade-off between a valid and an economical assessment. For instance, at one extreme one might implement a performance assessment potentially incorporating extensive observations by a rater. At the other extreme, a paper-and-pencil test limited to multiple choice items might be applied. Only when these questions are answered, the choice of the psychometric model will play its role.

This relativization of the psychometric approach must be put in perspective itself. While the impact of choosing a specific psychometric model on the educational practice may be indirect and limited, the consequences for the scientific quality of the assessment will be important. As mentioned above, models and data must fit together as much as possible or at least as necessary for the intended application.

NOTE

1. This article is written in the perspective of this group; members: Jean-Philippe Antonietti, Institut de mathématiques appliquées, Université de Lausanne, Jean Moreau, Unité de recherche pour le pilotage des systèmes pédagogiques du canton de Vaud, Lausanne, Urs Moser, Institut für Bildungsevaluation, Universität Zürich, Erich Ramseier, Bildungsplanung und Evaluation, Erziehungsdirektion des Kantons Bern.

REFERENCES

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? In E.V. Smith & R.M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 143-166). Maple Grove, MN: JAM Press.

Asendorpf, J.B. (1995). Persönlichkeitspsychologie: Das empirische Studium der individuellen Besonderheit aus spezieller und differentieller Perspektive. *Psychologische Rundschau, 46*, 235-247.

Brennan, R.L. (2006). Perspectives on the evolution and future of educational measurement. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1-16). Westport, CT: American Council on Education.

Brunner, M., & Süss, H.-M. (2007). Wie genau können kognitive Fähigkeiten gemessen werden? Die Unterscheidung von Gesamt- und Konstruktreliabilitäten in der Intelligenzdiagnostik für den Berliner Intelligenzstrukturtest. *Diagnostica, 53*, 184-193.

De Pietro, J.-F., Müller, R., & Wirthner, M. (2007). HarmoS-L1: vers des standards de base pour la langue de scolarisation - In Richtung Basisstandards im Bereich der Schulsprache. *Babylonia, 15* (4), 40-52.

EDK/CDIP (2005). *Harmonization of obligatory school in Switzerland: the most important projects. Facts - sheet.* Bern: Schweizerische Konferenz der kantonalen Erziehgunsdirektoren (EDK/CDIP).

Embretson, S.E., & Reise, S.P. (2000). *Item response theory*. Mahwah, NJ: Lawrence Erlbaum.

Griffin, P. (2007). The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation, 33*, 87-99.

HarmoS Konsortium Mathematik. (2007). *Kompetenzmodell HarmoS Mathematik. Papier 4.* Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK/CDIP).

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Expertise*. Bonn : Bundesministerium für Bildung und Forschung (BMBF).

Kolen, M.J. (2006). Scaling and norming. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155-186). Westport, CT : American Council on Education.

Kubinger, K.D., & Draxler, C. (2007). Probleme bei der Testkonstruktion nach dem Rasch-Modell. *Diagnostica, 51*, 131-142.

Mazzeo, J., Lazer, S., & Zieky, M.J. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 681-699). Westport, CT : American Council on Education.

National Research Council (2001). *Knowing what students know : The science and design of educational assessment*. Washington, DC : National Academy Press.

Network-A/INES/OECD (2004). *Review of assessment activities. Network A newsletter, issue 16*. Washington, DC : National Center for Education Statistics (NCES).

OECD (2005). *PISA 2003. Technical report*. Paris : OECD.

Ramseier, E., & Moreau, J. (2007). *Stichprobendesign und Gewichtung in der HarmoS-Validierungsstudie.* Bern : Schweizerische Konferenz der kantonalen Erziehguns-direktoren (EDK/CDIP).

Renaud, A. (2006). *Harmonisation de la scolarité obligatoire en Suisse (HarmoS). Design général de l'enquête et échantillon des écoles*. Neuchâtel : Office fédéral de la statistique.

Rost, J. (2004). Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. *Zeitschrift für Pädagogik, 50*, 662-678.

Scheerens, J., Glas, C., & Thomas, S.M. (2003). *Educational evaluation, assessment, and monitoring. A systemic approach.* Lisse : Swets & Zeitlinger.

Tristan, A. (2006). An adjustment for sample size in DIF analysis. *Rasch Measurement Transactions, 20*, 1070-1071.

Weinert, F.E. (2001). Concept of competence : A conceptual clarification. In D.S. Rychen & L.H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45-65). Göttingen : Hogrefe & Huber.

Wright, B.D., & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

Wu, M.L., Adams, R.J., Wilson, M.R., & Haldane, S. (2007). *ACER ConQuest version 2. Generalised item response modelling software.* Melbourne : ACER Press.

Yen, W.M., & Fitzpatrick, A.R. (2006). Item response theory. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, CT : American Council on Education.

Zieky, M., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement*. Princeton, NJ : Educational Testing Service.