PSICOPERSPECTIVAS
INDIVIDUO Y SOCIEDAD
VOL. 14, Nº 3, 2015
pp. 31 - 44

Validity and equity in educational measurement: The case of SIMCE

María Teresa Flórez Petour
Universidad de Chile, Chile
mtflorez@uchile.cl

ABSTRACT

Chile has more than two decades of assessing the quality of education through a national censal system called SIMCE. However, there are scarce studies that address this assessment system from an up-to-date concept of validity. The aim of this study was to critically analyse the validity of SIMCE, considering the unitary concept of validity, the argument-based approach, the consideration of the whole process involved in the assessment system, the inclusion of the interpretations of all the stakeholders involved in the process, and threats to validity. Interviews to key stakeholders were carried out along with analyses of official documents about SIMCE. An iterative analysis was executed on the basis of the theoretical considerations, focusing on the fairness of the assessment system as understood in international standards. Some of the main findings indicate SIMCE lacks fitness-for-purpose in relation to its aim of improving the equity of the education system. It generates negative consequences which can be detrimental to students with more learning difficulties and from more vulnerable groups. And, it does not consider in its construction the diversity of students who have to take the tests.

Key words: fairness; assessment; validity; SIMCE; standardised tests

Validez y equidad en la medición escolar: El caso de SIMCE

RESUMEN

Chile lleva más de dos décadas evaluando la calidad de la educación a través de un sistema nacional censal llamado SIMCE. Sin embargo, hay una escacez de estudios que se refieren a este sistema de evaluación desde un concepto actualizado de validez. El objetivo de este estudio fue analizar críticamente la validez del SIMCE, considerando el concepto unitario de validez, el enfoque basado en el argumento y la consideración de todo el proceso relacionado en el sistema de evaluación, la inclusión de interpretaciones de todos los interesados que están involucrados en el proceso y amenazas de validez. Se realizaron entrevistas a los principales interesados y se analizaron documentos oficiales del SIMCE. Se realizó un análisis de tipo iterativo, sobre la base de la teoría considerada, concentrándose en la equidad de los sistemas de evaluación como se entiende en estándares internacionales. Entre los principales resultados se encuentra que el SIMCE no cumple aquellos propósitos relacionados con una mayor equidad del sistema educativo; genera consecuencias negativas que podrían perjudicar a los estudiantes con más dificultades y de sectores más vulnerables; y no considera en su construcción la diversidad de estudiantes que deben responder las pruebas.

Palabras clave: equidad; evaluación; validez; SIMCE; pruebas estandarizadas

Como citar este artículo: Flórez Petour, M. T. (2015). Validity and equity in educational measurement: The case of SIMCE. Psicoperspectivas, *14*(3), 31-44. doi: 10.5027/PSICOPERSPECTIVAS-VOL14-ISSUE3-FULLTEXT-618

Recibido
2-03-2015

Aceptado
25-08-2015

The technical dimension that represents equity in an assessment process is fairness. According to the International Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014, p. 49), fairness is recognized as "a fundamental validity issue (which) requires attention throughout all stages of test development and use". The consideration of fairness requires test developers to control for aspects such as: standard administration conditions that assure equality of opportunities for all examinees to demonstrate their ability; attention to potential bias in score interpretation; difficulties in access to the construct as measured for some subgroups of the population; accuracy of score interpretation in relation to specific groups, among others (AERA, APA, NCME, 2014).

There is international consensus around the importance of the ethical dimension of testing since Messick's emergence in the field of assessment (1980, 1989). Additionally, literature from the field of sociology of education has highlightedhow testing systems affect specific minority groups in the assessed population(Filer, 2000). In this context, the scarcity of studies on these issues in relation to the Chilean national curriculum assessment system (SIMCE) seems surprising.

The Chilean System for the Measurement of Quality in Education (SIMCE) was established by law in 1990 (Ministry of Education, 1990) and has been in place for more than two decades. For several years it was under the control of the Ministry of Education through a specific unit focused on the test. In 2012, its development was transferred to the Agency for Quality in Education, a government institution independent from the Ministry.

SIMCE involves assessing the whole cohort of students of a specific level, including private, public and government subsidised public schools. It was initially given to students in the 4th and 8th levels of primary school and in the 2nd level of secondary school. Each year only one of these levels was assessed in the areas of Language and Mathematics, along with History and Natural Sciences for the first two levels (4th and 8th). After 2006, two levels were assessed annually: the 4th level was assessed every year along with one of the other two levels. After 2010 the areas of English as a Second Language and Physical Education were included in the system and the 2nd and 6th levels of primary school began to be assessed. This means that nowadays schools are exposed to around 17 tests per year.

Despite its long history and its increasing pervasiveness, only a few critical studies are found in the available literature. A critical trend emerged at the end of the 1990s but it mainly referred to the technical comparability of results from one year to the next.The controversy generated improvements through the introduction of IRT in test development, as well as equating mechanisms (Schiefelbein, 1998). Only one study focused on validity issuesaround SIMCE (Eyzaguirre& Fontaine, 1999), although it only analysed the dimensions of construct and content. Fairness was not addressed by these authors.

In light of the Ministry of Educactions's concerns and the critical attention of researchers around aspects such as the multiplicity of purposes of this assessment system (Comisión SIMCE, 2003;Bellei, 2002) the limited amount of studies around the validity of theinterpretations and actions based on SIMCE comes as a surprise. If studies on SIMCE in relation to validity are scarce, research that specifically focuses on its fairness is virtually inexistent. Despite this research gap, the assessment system is characterised by different sources as methodologically solid and as having a high level of credibility (Meckes& Carrasco, 2010; Comisión SIMCE, 2003, 2015), even though the corresponding evidence for these statements is not provided. Recent technical reports by the Agency for Quality refer mainly to validation processes in terms of content coverage and statistical procedures, while broader validity issues such as fitness-for-purpose or consequences are ignored (Agency for Quality in Education, 2014,2015).

A new wave of criticism has recently emerged in relation to SIMCE both from actors of the education system (Comisión SIMCE, 2015) and from research (Australian Council for Educational Research [ACER], 2013; Flórez, 2013; Ortiz, 2012). A new committee was appointed in 2014 to critically review the system and to propose ways in which its quality and functioning could be improved. The detailed conclusions of the committee are to be published in an extensive report during 2015 and an executive summary was already published in January 2015. It is, therefore, a favourable moment to contribute to this discussion by suggesting potential improvements to validity and fairness in SIMCE. Given that one of the stated purposes of SIMCE, according to several official documents, is to improve the equity of our education system, fairness becomes a particularly salient issue in this discussion.

This article is based on a wider project that was funded by the National Council for Education in 2012-2013 and supported by the Oxford University Centre for Educational Assessment (OUCEA). In this project, the author critically analysed the validity of interpretations and actions related to SIMCE, considering a qualitative approach where both interviews with key stakeholders and official documents on SIMCE were used a as a source for analysis.

Findings are organised following some of the categories developed by Stobart (2009) in relation to validity: (i) purposes and fitness-for-purpose; (ii) administrative reliability; (iii) test construction and interpretation of results; (iv) impact/consequences.For each aspect a brief theoretical explanation is provided, along with threats to validity as defined by Stobart (2009).

In general terms, the perspective on validity that permeates this paper is that of the unitary concept developed by SammuelMessick (1980, 1989), in which all dimensions (construct, content, criterion, consequences) are subsumed into the assessment construct, and where validity is defined as "(…) an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness of inferences and actions* based on test scores or other modes of assessment" (Messick, 1989, p. 13).

Following later developments of the concept, this paper adopts an argument-based approach to validity (Kane, 2008, 2010, 2011). It also considers that the whole assessment process is important (Newton, 2013), including the perspectives and interpretations of different stakeholders involved in the system (Koch&DeLuca, 2012; Hubley&-Zumbo, 2012). Attention is also given to consequences as an essential aspect in validity studies (Kane, 2008, 2010, 2011; Messick, 1980, 1989; Shepard, 1997; Linn, 1997).

Method

Methological framework and research questions

Due to the current predominance of a pragmatic and operational approach to validity favoured in the work of assessment development agencies, qualitative and critical approaches are seldom explored in relation to specific assessment systems. Therefore, there is a general scarcity of evidence in relation to aspects such as: stakeholder's interpretations and views in relation to test results and around its quality and fitness-for-purpose; the validity of interpretations provided in official sources; the quality of questions and, therefore, their suitability as a means to cover a specific domain; the quality of the processes involved in the development and implementation of SIMCE, going beyond their mere description; among other aspects relevant to the construction of a validity argument.

A qualitative approach was adopted in the study as a means to begin filling this research gap by exploring the kinds of evidence thatthis perspective couldcontribute in the construction ofan argument in relation to the validity of inferences around SIMCE.With this approach as a basis, the main research question that guided the study was:Is SIMCE a valid assessment process, considering the different dimensions of the concept of validity?

Decisions about the types of evidence to be collected were made on the basis of recent developments in validity theory. Considering the unitary concept and the argument-based approach to validity, as well as Stobart's (2009) dimensions of validity, three specific sub-questions were defined: a) Is SIMCE valid for all the purposes and uses currently attributed to it?; b) What are the constructs assessed in SIMCE? Is there a consistent interpretation about these constructs throughout the system?; and c) What contents does SIMCE assess and how representative is that content of the construct to be assessed?

Along with these aspects, over the last few years, the international literature has insisted on the importance of interpretations from the perspectives of a broad range of stakeholders involved in the assessment process, as well as onthe need to consider all the stages of the assessment system (Newton, 2013; Koch & DeLuca, 2012). Therefore, rather than a focus on the instrument and its internal characteristics -a perspective that has been advocated by authors such as Lissitz&Sammuelsen (2007) and has been widely criticised by the main voices of the field (see for example Sireci, 2007; Kane, 2008; Moss, 2007)-, this study is consistent with current approaches to validity, which are broad in their consideration of both the agents and the phases of the assessment process.

The views and interpretations of test developers, coordinators and internal professionals in SIMCE, item constructors, open question reviewers, teachers, among other actors, were included. On the one hand, this allowed evaluating whether their interpretations around the construct were consistent throughout the process. On the other hand, these sources served as a means to generate data about the quality of each stage of the process and, therefore, contributed to the overall judgement on the quality of this assessment system as whole. 3 additional sub-questions guided this dimension of the study: a) What do different actors of the process think about the validity of SIMCE?, b) Is there a common view among actors about the validity of SIMCE and its results?, and c) Is validity maintained throughout SIMCE's processes of production, distribution, correction and use?

Data collection and analysis

Two sources of data were considered in the study:

Semi-structured interviews

Starting from previous working contacts of the researcher, snowball sampling was used to gain access to different stakeholders who have been involved in the process of SIMCE in a variety of roles. Interviews were carried out with 15 stakeholders involved in areas as varied as coordination, item construction, supervision of item construction, item reviewing, supervision of the reviewing process, construction and validation of criteria for open questions, with some of them having held more than one role in their career. These actors work in the areas of Spanish and Mathematics (unless they hold a general more overarching position) and are both internal and external to SIMCE. Their years of involvement allow for obtaining perspectives about the test from 1998 to 2013 and, therefore, include participants that have worked for SIMCE under different political administrations.

Table 1 provides more details about these participants, including their roles and disciplines (when applicable). However, due to anonymity issues their years of involvement are not stated and their roles are made more general in some cases. Being a policy elite group, the value of the data lies not in the number of participants but in a type of knowledge that is usually restricted, despite this being the place where the initial interpretations around SIMCE are generated. The value increases in the case of actors internal to SIMCE, in terms of their privileged access to the test and its procedures.

The views of teachers are seldom included in arguments for validation, although they constitute one of the main recipients of the results of the national test and despite the existence of evidence from research and policy reports about their lack of understanding and use of the scores of SIMCE (Comisión SIMCE, 2003; Centro de Investigación y Desarrollo de la Educación [CIDE], 2007; Taut,Cortés, Sebastian&Preiss, 2009). Due to this gap,interviews were carried out with 20 teachers in order to explore their perspectives in relation to the interpretation, use and consequences of the results.

Participants were selected through a purposeful sampling procedure that aimed at a wide variety of contexts (rural/urban, public/private/government subsidised private), levels of experience (10 recently graduated teachers and 10 experienced teachers), disciplines (included and not included in SIMCE), gender and type of initial teacher education. The aim of variety was not generalisation but discursive representativeness in qualitative terms. However,although it was a small sample, the degree of consistency in the views of teachers around SIMCE was high and was also consistent with views expressed through research of a more quantitative nature (CIDE, 2012; Taut et al., 2009). Table 2 provides details of the profiles of each participanting teacher.

Ethical requirements were met by informing the participants about the project and securing their anonymity (all the names used here are pseudonyms). All of the participants signed consent forms.

After transcription, interviews wereanalysed using NVivo 10, in an iterative process of content analysis (Srivastava& Hopwood, 2009), which had the theory around validity as a basis. This analysis involved exploring aspects such as: (a) purposes of SIMCE; (b) description of their involvement in the process and their perceptions about its quality; (c) the construct assessed by SIMCE; (d) the means to represent such construct; (e) their knowledge and views about uses, interpretations and consequences of SIMCE; (f) and their perceptions about the general validity of the system.

Documents

Clarity around the purposes and interpretations of scores and results is a crucial aspect of validity(AERA, APA, NCME, 2014).Official documents provided by the assessment agency that creates the tests are, therefore, a fundamental source in the construction of an argument around validity. A set of publicly available documents was selected with this aim in mind: to observe how the purposes, uses and interpretations of the scores of SIMCE are officially defined. A second group of documents was selected following the same purpose asthe study of Eyzaguirreand Fontaine (1999): to analyse the construct and content assessed by SIMCE through its questions. In order to make data more manageable, these documents were restricted to the areas of Spanish and Mathematics (the full list of selected documents can be found in Appendix 1).

Shorter documents (sections of websites and brochures with general guidelines) were coded in NVivo using the same inductive and iterative approach (Srivastava & Hopwood, 2009). Initially,the 373 codes induced from the data were later classified in fewer codes. This analysis served as a basis for the design of the interviews. In a second phase, more extensive documents (results reports, explanation of levels of achievement, etc.) were analysed

Table 1.
Selected key actors in the process of SIMCE

	Pseudonym	Role(s) in SIMCE	Discipline (if applicable)
1	Alicia	Internal SIMCE professional / Coordination role	Mathematics
2	Alejandra	Supervisor of item construction Supervisor of item reviewing	Language
3	Arturo	Internal SIMCE professional / Coordination role	Not applicable
4	Augusto	Item reviewer	Mathematics
5	Daniela	Item constructor Item reviewer Validator of assessment criteria rubrics	Language
6	Emilio	Item reviewer	Mathematics
7	Jaime	Internal SIMCE professional / Coordination role	Mathematics
8	Josefina	Internal SIMCE professional / Coordination role	Not applicable
9	Pedro	Institutional coordinator of item construction (external university hired for the process)	Language
10	Rebeca	Internal SIMCE professional / Coordination role	Language
11	Roberta	Supervisor of item construction	Language
12	Rodolfo	Item reviewer	Language
13	Sandra	Supervisor of item correction Item reviewer	Language
14	Sofía	Internal SIMCE professional / Coordination role	Language
15	Ximena	Supervisor of item correction Item reviewer	Language

through pen and paper coding, using the codes induced-from the short documents as a reference.

A group of experts was consulted during the research process as a means of avoiding potential bias in data analysis. Two assessment experts from OUCEA provided feedback on the interview schedules, which were reviewed and modified based on their critiques. One of the interviewees with expertise in Mathematics was consulted in the process of the analysis of items. Finally, concerns about the validity of SIMCE were discussed with an expert psychometrician in order to confirm or refute their relevance.

Results

In this section, the results of this study deemed relevant to equity issues in SIMCE are described in detail. Findings are organised considering four sections, taking Stobart's (2009) categories as a reference: (a) purposes and fitness-for-purpose; (b) administrative reliability; (c) test construction and interpretation of results, and (d) impact/consequences. When necessary, a brief theoretical introduction to validity issues is provided at the beginning of the section, followed by research results. For reasons of space, the author provides only a few illustrative examples of the analysed sources but these are representative of major trends in data (further evidence can be found in Flórez, 2013).

Purpose and fitness-for-purpose

Despite its historical development, there is long-standing agreement in the literature on validity that a test is not valid in itself. That is, the test is not valid as such in any given context but the inferences that are drawn from it on the basis of a specific purpose and/or use (Cronbach, 1984; Anastasi&Urbina, 1997; Kane, 2011; Newton 2012; Hubley & Zumbo, 2011). Cronbach (1984), for instance, states that the question is not how valid this assessment is

Table 2.

Characteristics of participant teachers

Profile	Pseudonym	Level of teaching	Gender	Type of school	Type of school location	Discipline (if specialised)
Recently graduated	Rosaura	Secondary	F	Private GS	Urban	English as a foreign language
	Catalina	Primary	F	Public	Urban	English as a foreign language
	Saúl	Primary	M	Private GS	Urban	History, Geography and Social Sciences
	Ernesto	Primary	M	Public	Rural	Not applicable
	Susana	Secondary	F	Private GS	Rural	Physical education
	Raquel	Secondary	F	Private NGS	Urban	Philosophy
	Ana	Secondary	F	Public	Urban	Biology
	Óscar	Primary	M	Public	Urban	Spanish
	Felipe	Secondary	M	Public	Urban	History, Geography and Social Sciences
	Patricia	Secondary	F	Public	Urban	Biology
Experienced	Leila	Secondary	F	Public	Urban	Mathematics
	Néstor	Secondary	M	Public	Urban	Philosophy
	Rosa	Primary	F	Private GS	Urban	Spanish
	Hernán	Secondary	M	Private NGS	Urban	History, Geography and Social Sciences
	Amelia	Primary	F	Public	Rural	Spanish
	Fabiana	Primary	F	Public	Urban	Natural Sciences
	Luisa	Primary	F	Public	Urban	History, Geography and Social Sciences
	Marcos	Primary	M	Private GS	Urban	Musical Education
	María	Secondary	F	Public	Urban	Spanish
	Laura	Primary	F	Public	Urban	Technologic Education

*Note:*GS: Private government-subsidised schools; NGS: Private non-goverment-subsidised schools

but what decisions is it valid for, which is consistent with Messick's idea of the adequacy of actions and inferences derived from the results of a test. Therefore, one central aspect in the analysis of the validity of an assessment system has to do with its purposes and intended or unintended uses, as well as the interpretations that are derived in connection to each of these purposes. In Stobart's framework, some potential threats to validity in relation to the purposes of an assessment system are: "lack of clarity; competing purposes; unachievable purposes'. In terms of fitness-for-purpose, the author points out that a central question would be: "Does the assessment do what it is claiming to do?" (Stobart, 2009, p. 165).

One of the most widely mentioned and long-standing purposes of SIMCE, as indicated in the analysed documents as well as in the very recent executive summary of the 2014 Committee, is to improve the quality and equity of the education system.A central question about validity and equity in this respect would be: Is SIMCE increasing the equity of the education system?

One of the main findings of the study presented here is that SIMCE has, according to interviewees and documents, at least 17 different stated purposes (2 of them are not stated in official documents but are mentioned by interviewees), along with two macro-purposes in tension: the pedagogical use of data (here results are understood as a support for teachers' reflexive processes around their practice and as a means to improve their teaching), and holding schools accountableby exerting pressure on them. This problem is recognised by internal members of SIMCE, as can be seen in this quotation from Arturo:

'What happens is, when you have a system with too many purposes, finally that, literature is very extensive on that I think, tensions emerge (…). My impression is, in that context, SIMCE, in a general sense, has probably satisfied in a better way the accountability demands and the ones that have to do with political monitoring of practices, rather than [the ones related to] feedback on practice, historically speaking. So, despite the efforts, I still think there is some unbalance in how well all these purposes are satisfied. (…) when one sees the impact on the system, one realises there are purposes (…) that are better accomplished than others. In that sense, there is clearly (…), I don't know if a tension, but when decision-making comes, I think the system satisfies better its purposes of monitoring and accountability than the purpose of feedback on practice. (…) we haven't found formulae to make the message we send to headteachers and teachers more powerful' (Paragraph 18).

One assumption that underlies the purpose of improving equity in education is that the assessment system will put pressure on schools and thus motivate them to increase the quality of the service they are providing, and therefore more students from different backgrounds will have access to a better education. Additionally, the macro-purpose associated with pedagogical use presupposes that teachers receive information that allows them to improve their practice and thus the learning of their students, generating as a result the provision of quality education for more students. Both assumptions are discussed below.

Data for accountability

Pressure is exerted through the publication of the results of SIMCE in a context where schools have to compete with each other to attract more students in a voucher system model. Private, public and government-subsidised schools have to participate in this competition as if they were equal, despite recognition that the Chilean education system is highly segregated (Valenzuela,Bellei& De los Ríos, 2010; Núñez, 2015). Public schools in Chile are the ones that receive the most vulnerable students, especially those in which selection is not permitted. These are also the schools that usually obtain the lowest test scores. As a result, public schools suffer a decrease in student intake, which could potentially lead to their closure, and are portrayed as deficient through the media blitz that is generated each year around the scores of SIMCE. International standards on fairness indicate that "(…) potential inequities in school resources available to students from traditionally disadvantaged groups (…) affect the quality of education received. To the extent that if inequity exists, the validity of inferences about student ability drawn from achievement test scores may be compromised"(AERA, APA, NCME, 2014, pp. 56-57)

Therefore, SIMCE does not consider gaps in the opportunity to learn that different types of schools are able to generate. Although official reports indicate schools should only be compared to other similar schools, in practice and in the broader context of a market-centred model of education, this is not taken into account. Teachers interviewed in this study criticise not the existence of a national assessment system but the lack of fairness of SIMCE as part of a market-oriented school system. Hernán, for example, who works at a private school where they supposedly do not care about SIMCE results, indicates:

'There is a certain amount of restlessness too, I mean, we say it's not important, nothing standardised, but there is always a concern (…). And we always finally end up validating that this shows pedagogical level, the academic level of our students, educational level… (…). But is very distorting, because eventually you are in a school that supposedly doesn't care, like this, but finally here they also do preparation tests for SIMCE. (…) And you know why? Because it has to do with the market. Because if [name of the school] falls from 290 points (…) or it's not near 300, less students will enter here' (Paragraph 93).

If the publication of results tends to stigmatise vulnerable groups, it is unlikely that the purpose of improving the equity of the system is being achieved. It seems the result is rather the opposite. Ortiz (2012) adds to this that another purpose of SIMCE, that is, its original purpose of informing parents' school choice, has only been achieved for those who can pay to attend schools with better results, which does not contribute to the equity of the system. According to Ortiz (2012), it has also contributed to a weaker image of public schools by stigmatising them, although research has shown it is not the effect of the school that influences results but mainly the socio-economic background of parents.

Pedagogical use of data

The second assumption draws on the pedagogical use of data as a macro-purpose of SIMCE. There are two studies with consistent results around the use that parents and teachers make of data derived from the test. The series of reports from CIDE (2007, 2008a, 2008b) andSepúlveda(2008) conclude that teachers value the information that is provided, especially after the introduction of achievement level descriptions in 2006. However, they do not see the same value in the comparisons that are established in reports and they seldom use data from SIMCE to

establish future goals, actions or commitments. Likewise, Taut et al. (2009) conclude that both parents and teachers experience problems to recall and interpret basic information from the reports correctly. The developers of SIMCE have not provided evidence that contradicts these results and there is no empirical basis to state that this assessment system has contributed to increase the quality and equity of education in Chile. In the context of a Messickean and argument-centred approach to validity, providing this evidence would be crucial to ensure both the validity and the fairness of the test.

Neither pressure nor the pedagogical use of data are effective means to improve the equity of the system. This lack of impact in equity is increased by the simultaneity of both purposes, as the former seems to make the second less feasible. Literature recognises the difficulties of expecting the same test to support both purposes. Pressure has consequences that do not necessarily improve teaching quality and even test construction aspects vary widely depending on the purpose that is chosen (Haldane, 2009). More recent trends are moving towards a focus on one of the two purposes (see for example Moe, 2009 for the case of Norway; Wandall, 2009 for the case of Denmark; and the System of Assessment of Learning website (SEA, 2015) developed by the Council for Initial and Primary Education and the Department for Assessment of Learning in Uruguay) or designing a system of multiple assessments where both purposes are addressed through different instruments and with an emphasis on higher order thinking skills as well as high quality teaching (see for example Herman and Linn, 2013 for the case of the Smarter Balanced Assessment Consortium in the US).

Administrative reliability and fairness in SIMCE

Current assessment standards recognise reliability and validity as two closely interlinked aspects of test development (AERA, APA, NCME, 2014). Stobart (2009) considers reliability as one component of validity, which can be threatened by factors such as: "Security breaches; inconsistent test administration and conditions; inappropriate modifications/time constraints; test-taker reliability" (Stobart, 2009, p. 165).Along the same line, standards recognise that a fundamental aspect of fairness in testing is related to administration and scoring procedures that minimise construct-irrelevant variance and promote valid score interpretations for all examinees (AERA, APA, NCME, 2014). Some concerns emerge from the results of the study presented here in relation to potential inconsistencies in test administration and conditions as well as test-taker reliability.

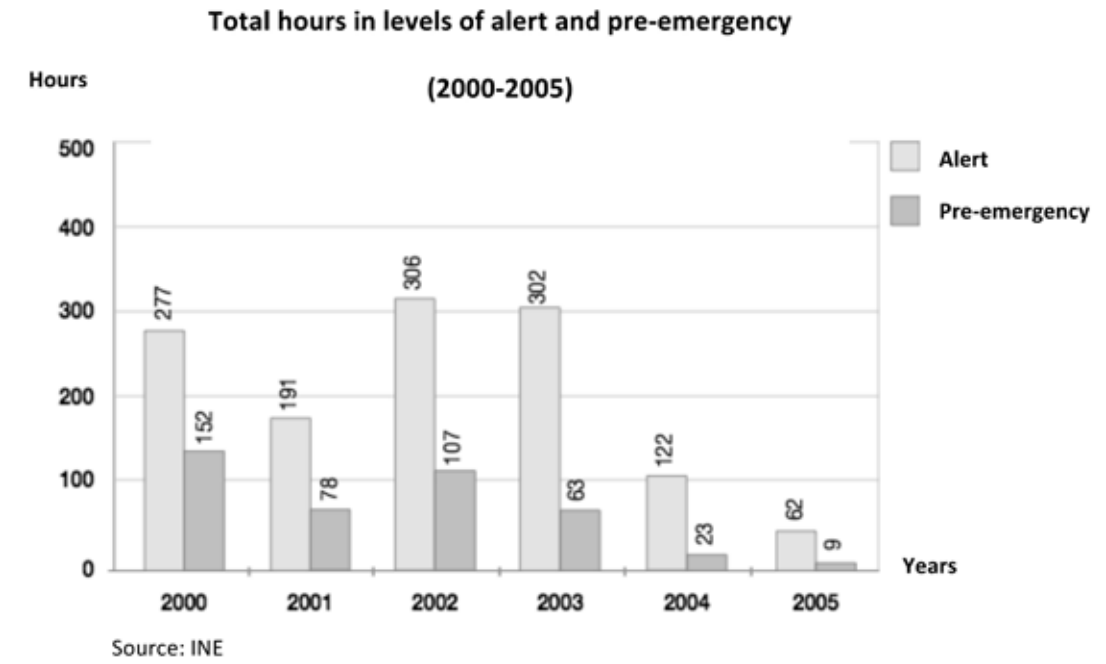
Teachers and some members of SIMCE recognise in their interviews a series of malpractices oriented to artificially increase the score of the school in the test. One of these practices is to either ask lower performing students not to attend the school on the day of the test or to provide high performers with special training in order to boost the score of the whole group through their performance. This is preliminary qualitative evidence on which further research must be carried out. If these effects are widespread, they could not only affect the validity of score interpretations as the administration of the test would not be homogeneous, but also have serious implications in terms of fairness. The self-esteem of students with low performance levels could be damaged as well as their motivation to learn. Unlike their high-performing peers, they would also have less opportunity to learn and thus less access to the construct being assessed. It is imperative, therefore, to carry out studies both on the reach of these types of consequences and on the potential washback effects they might have on score interpretation.

Test construction and result interpretation

According to interviewees from SIMCE and to the recent technical report on the 2012 process (Agency for Quality in Education, 2014), SIMCE considerspotential gender and rural/urban bias in their piloting processes, and eliminates or modifies questions that present this type of issues. However, they do not include socio-economic bias in the post-pilot analysis of items. Given the consequences of the test for public schools and the persistence of socially segregated results every year, to consider this potential source of bias would be essential. As Standard 3.3 points out (AERA, APA, NCME, 2014, p. 64): "Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test".

The example below was taken from a 2010 document aimed at illustrating questions associated with different levels of achievement in 8th grade Mathematics. It illuminates how SIMCE questions could involve socio-economic bias:

Alert and pre-emergency correspond to terms used in the context of a policy against pollution that operates only in a few contaminated cities in Chile, mainly the capital Santiago and Temuco in the south of the country. It categorises different degrees of pollution, according to which the circulation of cars is prohibited to a larger or lesser extent, using the last number of the car's license plate as a reference.



How many hours of difference were there between the pre-emergencies of 2003 and 2005?

A. 54
B. 60
C. 63
D. 72

Figure 1. Example of question associated with different levels of achievement in 8th grade Mathematics

As can be seen in the example, no context is provided in the question. As a result, students whose parents have a car and live in Santiago or Temuco could have an advantage over students who live in other cities and/or do not own a car. As rural/urban bias is controlled for in the post-pilot psychometric analysis, the regional aspect might be covered. However, there is no control for socio-economic bias in relation to families who use public transport instead of a car and, therefore, for whom the context of the item would be less familiar. This type of bias should also be considered, for example, in terms of topics, cultural references, language, etc., both in stimuli and questions. As Standard 3.1(AERA, APA, NCME, pp. 63-64) indicates: "Test developers need to be knowledgeable about group differences that may interfere with the precision of scores and the validity of test score inferences, and they need to be able to take steps to reduce bias".

Another example that is repeatedly provided in the section on fairness of the Standards(AERA, APA, NCME, 2014) is that of students whose native tongue is different from the language of the test. The two Mathematics open question reviewers interviewed in the study were consistent in recalling a problematic situation in this respect. Emilio, a Mathematics item reviewer, describes it in the following way:

> '(…) we received an answer in a strange language (…); we assumed it was an indigenous language, mapudungún, (…) within this answer, which was illegible to us, (…) you could perceive there was something going on but we were not able to understand it and that was immediately coded as wrong, that is, there wasn't any process to find out what could be going on there but [they just said] <<no, this is illegible, is wrong>>, and I didn't like that. That happened to me and I heard of similar cases a couple of times and that seemed strange to me. At least I would have set it aside and investigated a bit (…)' (Paragraph 17).

According to the 2012 census, more than 10% of the Chilean population belongs to an indigenous group (Institu-

to Nacional de Estadística [INE], 2012). Despite this cultural variety, there are children in Chilean schools whose native tongue is different from Spanish who have to take the SIMCE test in a language that could potentially be more difficult for them than for other students. The excerpt shows how the student felt more confident providing an answer in Mathematics in his/her own language. It also reveals that the student's answer might have been correct but it was discarded only on the basis of being written in a language unknown to the reviewers. This has implications in terms of fairness, as language in this case is a barrier for examinees and thus a potential source of construct-irrelevant variance.

Another related aspect that is mentioned in the Standards(AERA, APA, NCME, 2014) is the suggestion to carry out sensitivity reviews of tests in order to determine whether indigenous groups might have alternate interpretations of particular questions due to issues more sensitive to them. There is no information about the way SIMCE currently addresses this issue.

It is likely that due to these gaps teachers perceive SIMCE as a test that treats very different schools and contexts as if they were homogeneous, making the information provided by the test not very useful for their specific circumstances and unfair in the sense that the same demands are placed on schools that do not have the same resources and conditions. Ernesto, who is a primary teacher in a public rural school, explains this issue in the excerpt below, which illustrates the general perception of teachers in this study:

> '(…) SIMCE ties my hands because its what, its almost the life project of the whole school. Schools live and die for SIMCE and there is [] an overvalued importance due to, due to the fact that there are no internal assessments in the country that give [], that are able to replace SIMCE or of generating [] a panorama of Chilean education with more precision [] or at least more meaningful in relation to all the different worlds that are found in this Chile, because SIMCE is horizontal and affects equally and measures equally the knowledge of all the realities of students, of the country's pupils. Which is, is a stupid thing' (Paragraph 32).

These perceptions are consistent with the findings from a large-scale survey carried out each year by CIDE (2012), where different actors of the school system respond to a series of questions, including a separate section about SIMCE. The 2012 version of this survey shows that 88,1% of teachers and 87,2% of headteachers agree that "SIMCE, by comparing schools that are very different in the socioeconomic and cultural features of their students, ignores the effort made in context with more difficult areas" (CIDE, 2012, p. 44). Along the same lines, teachers demand in their interviews a more context-sensitive assessment system, where interpretations and intended consequences take these differences into account.

The reports of results that were analysed in this study are careful in saying that schools should only be compared to those of a similar type. However, national reports present the scores of different types of schools and from different socio-economic backgrounds in the same table, which could motivate non-valid (and unfair) comparisons. Additionally, policies tied to SIMCE results provide material resources to schools in proportion to their increase in results.They also involve sanctions to schools who repeatedly fail. These policies treat all schools as if they were all the same and were competing for resources in the same conditions.

Impact and consequences

Ever since Messick's (1980, 1989) development of his unitary concept, and after a long-standing debate,the consequences of a test were finally incorporated in the AERA, APA, NCMEStandards in 1999 as part of the validity of an assessment system.In relation to this aspect, Stobart (2009, p. 165) indicates as potential threats: "limited confidence in results; contested interpretations of results; inappropriate decisions made on the basis of results; negative impact on teaching and learning".All the teachers in the study recognised SIMCE as a pressure device that generates negative feelings in them as well as a series of negative consequences.

Teachers from the assessed disciplines, including English and Physical Education, refer to experiences of role conflict (Berryhill, Linney & Fromewick, 2009) caused by the test. That is, their concept and principles around pedagogy and around the discipline they teach conflict with the demands of SIMCE. Ana, a young Biology teacher at a public school for girls, portrays this experience in the excerpt below:

> '(...) the school in one way or another moves you to work aiming at (…) SIMCE, in some cases. Because finally, from years 9 to 12, what do you do? You acquire the practice of privileging multiple choice questions and that the girls learn to solve them, or at least they have the tools to solve or eliminate some, [] why this option

> is not the right one, why number one is not there and you have to restrict yourself to that. (...) But how real is that in terms of my teaching practice, in terms of what I am doing, there is an ethical issue there. (...) And you see yourself forced in one way or another by the system, because what do they do to us, that there is a bonus because of SIMCE, (…) then you privilege in some cases, more than education itself, to work aiming at those bonuses, and the system pushes you to work this way' (Paragraphs 78 to 86).

Additionally, these teachers declare feeling pressured to prioritise curriculum coverage instead of depth in learning.

In contrast, teachers in disciplines that are not assessed by the testperceive that their areas are overshadowed by what Néstor denominates 'the star disciplines: Spanish and Mathematics' (Paragraph 56). In practice, according to interviewees, this involves using hours of these areas for training or reinforcement of SIMCE; encouraging these teachers to plan their disciplines in a way that is functional to the assessed areas (mainly Spanish in this case); and the concentration of the material and human resources of the school in the levels and disciplines to be assessed each year.

Consistent with the results of Taut et al. (2009), all participant teachers recognise the existence in their working contexts of some form of preparation for SIMCE. The majority of teachers in this study, along with some of the interviewees from SIMCE, also refer to a series of malpractices that have resulted from SIMCE, such as: economic or material incentives for teachers and students; academic incentives for students; special training of high-performing students during the hours dedicated to areas not assessed by SIMCE as a means to boost the score; and replacement of late afternoon extra-curricular optional workshops for training and reinforcement of SIMCE.

This evidence shows how SIMCE distorts the practices of schools and how teaching and learning in a more holistic and meaningful sense is turned into artificial training situations focused on the areas that are assessed in the test, and on the type of questions that predominate in it. The results from the CIDE survey (2012, p. 44) also confirm this perception, as almost 60% of teachers agree that "SIMCE generates an increase in the mechanisation of teaching and learning and it impoverishes education". No difference in results was found in the author's study in terms of type of school in this respect, so an initial interpretation could be that SIMCE is equally negative for all Chilean students, as it impoverishes education across all social groups. However, given the importance of socioeconomic background in Chilean education, high-class students have more opportunities for a balance of this impoverishment through the economic and cultural capital of their families, a possibility that students from more vulnerable backgrounds would not have. The negative consequences of SIMCE are, therefore, more pervasive and inevitable for students from vulnerable backgrounds who also attend public schools, where the pressure for improvement is higher than in other contexts.

Conclusions

The results presented in this paper illustrate that validity and fairness are two intertwined areas in the development of a good quality assessment system. The former is centred on the adequacy of inferences derived from test scores, while the latter answers the question of whether these interpretations are valid for all examinees in terms of eliminating potential bias or systematic disadvantages for specific groups in the assessed population. In the case of SIMCE, the results of thisvalidity study revealed the following concerns around fairness:

(i) No evidence is found of the test as being able to fulfil its purpose of improving the equity of the education system.

(ii) There are potential reliability issues that demand further research to demonstrate that test administration is being fair for all students.

(iii) No steps are taken in test development in order to control for socio-economic bias or language and cultural barriers as potential sources of construct-irrelevant variance.

(iv) Further research is needed to evaluate how widespread the negative consequences of SIMCE are, not only in terms of general washback effects of these consequences on the interpretation of scores, but also of potentially increased detrimental effects for more vulnerable schools.

The distance between the concerns raised by assessment theory, which are reflected in international standards(AERA, APA, NCME, 2014), and what assessment agencies do in practice has been subject of long-standing criticism (Haertel, 1999; Anastasi, 1986; Hubley&Zumbo, 2011; Koch & DeLuca, 2012).

Validity seems to be reduced in many cases to an instrumental and pragmatist approach wherecontent coverage and statistical validation are considered to be sufficient to affirm that an assessment system is technically solid - a statement commonly found among advocates of SIMCE.As Kane (2010) recognises, following currentstandards and theoretical agreements would make validation a much more complex process. However, this is no reason to avoid necessary changes. When the consequences of a test are examined in connection to equity issues in testing, as illustrated in this paper, then this complex approach becomes an ethical imperative.

References

ACER (2013).*Evaluation of the processes and products related to the production of instruments, field operations and data management of national SIMCE tests.* Santiago de Chile: ACER and Agency for Quality in Education.

AERA, APA, NCME (2014). *Standards for educational and psychological testing*. Washington D.C: AERA.

Agency for Quality in Education (2014).*InformeTécnico SIMCE 2012* [SIMCE Technical Report 2012]. Retrieved fromhttps://s3.amazonaws.com/archivos.agenciaeducacion.cl/documentos-web/Informe_Tecnico_Simce_2012.pdf

Agency for Quality in Education (2015). *Informe Técnico SIMCE 2013* [SIMCE TechnicalReport 2013]. Retrieved from:

http://archivos.agenciaeducacion.cl/documentos-web/InformeTecnicoSimce_2013.pdf

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37,* 1-15.doi: 10.1146/annurev.ps.37.020186.000245

Anastasi, A &Urbina, S. (1997). *Psychological testing*. Nueva York: Prentice-Hall.

Bellei, C. (2002). *Apuntesparadebatir el aporte del SIMCE al mejoramiento de la educaciónchilena* [Notes to debate on the contribution of SIMCE to the improvement of Chilean education].Disponible en http://biblioteca.uahurtado.cl/ujah/reduc/pdf/pdf/9204.pdf

Berryhill, J., Linney, J.A. & Fromewick, J. (2009). The effects of education accountability on teachers: Are policies to-stress provoking for their own good? *International Journal of Education Policy and Leadership, 4*(5), 1-14.

Centro de Investigación y Desarrollo de la Educación (2007).*Informe Final EstudioExploratorioEntrega de Resultados SIMCE con Niveles de LogroaEstablecimientosEducacionales Durante el año 2007* [Final report Exploratory study on Provision of SIMCE results with Levels of Achievement to schools during 2007]. Santiago de Chile:Centro de Investigación y Desarrollo de la Educación,Universidad Alberto Hurtado.

Centro de Investigación y Desarrollo de la Educación (2008a). *Informe cualitativo estudio Evaluación de la jornada de análisis de resultados SIMCE 2007* [Qualitative report study on the analysis of the results of SIMCE 2007 conference]. Santiago de Chile: Centro de Investigación y Desarrollo de la Educación, Universidad Alberto Hurtado.

Centro de Investigación y Desarrollo de la Educación (2008b). *Informe cuantitativo estudio "Evaluación de la jornada de análisis de resultados SIMCE 2007"*[Quantitative report study on the analysis of the results of SIMCE 2007 conference].Centro de Investigación y Desarrollo de la Educación, Santiago de Chile: Universidad Alberto Hurtado.

Centro de Investigación y Desarrollo de la Educación (2012). *IX Encuesta a Actores del Sistema Educativo 2012* [IX SurveytoActors of theEducationSystem]. Santiago de Chile: Centro de Investigación y Desarrollo de la Educación, Universidad Alberto Hurtado.

Comisión SIMCE (2003). *Evaluación de aprendizajes para una educación de calidad* [Assessment of learning for an education of quality]. Retrieved from

http://www.agenciaeducacion.cl/wp-content/uploads/2013/02/Comision_Simce.pdf

Comisión SIMCE (2015). *Informe Ejecutivo Equipo de Tarea para la Revisión del Sistema Nacional de Evaluación de Aprendizajes* [Executive Report of the Team for the Task of Reviewing the National System of Assessment of Learning]. Retrieved from

http://www.mineduc.cl/usuarios/mineduc/doc/201502021034480.InformeEjecutivoEquipoTarea.pdf

Council for Initial and Primary Education and the Department for Assessment of Learning of the Research, Assessment and Statistics Unit Division (2015).System for the Assessment of Learning in Uruguay [Sistema de Evaluación de Aprendizajes, SEA, Uruguay]. Retrieved from http://www.anep.edu.uy/sea/?page_id=2542

Cronbach, L. J. (1984). *Essentials of psychological testing*. New York: Harper & Row.

Eyzaguirre, B. & Fontaine, L. (1999). ¿Qué mide realmente el SIMCE? [What does SIMCE actually measures?]. *EstudiosPúblicos*, *75*, 107-161.

Filer, A. (2000) (Ed.).*Assessment. Social Practice and Social Product*. London: Routledge. doi: 10.4324/9780203465844

Flórez, T. (2013). *Análisiscrítico de la validez del SIMCE* [Critical Analysis of the Validity of SIMCE]. Retrieved from

http://www.cned.cl/public/Secciones/SeccionInvestigacion/investigacion_estudios_documentos.aspx

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, *18*(4), 5-9. doi: 10.1111/j.1745-3992.1999.tb00276.x

Haldane, S. (2009). Delivery platforms for national and international computer-based surveys history, issues and current status. In F. Scheuermann&J. Björnsson (Eds.), *The Transition to Computer-Based Assessment. New Approaches to Skills Assessment and Implications for Large-scale Testing* (pp. 63-67). Luxembourg: European Commission, Joint Research Centre Institute for the Protection and Security of the Citizen.

Herman, J. & Linn, R. (2013). *On the Road to Assessing Deeper Learning: The Status of Smarter Balanced and PARCC Assessment Consortia*. California: CRESST/University of California.

Hubley, A. &Zumbo, B. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, *103*(2), 219-230. doi: 10.1007/s11205-011-9843-4

Instituto Nacional de Estadística (2012). *Síntesis de Resultados Censo 2012* [Summary of Results. 2012 Census]. Santiago de Chile: Instituto Nacional de Estadística. Retrievedfrom

http://www.iab.cl/wp-content/themes/IAB/download.php?archivo=11803%7Cresumencenso_2012.pdf

Kane, M. (2008). Terminology, emphasis, and utility in validation.*Educational Researcher*, *37*(2), 76-82.doi: 10.3102/0013189X08315390

Kane, M. (2010). Validity and fairness.*Language testing*, *27*(2), 177-182. doi: 10.1177/0265532209349467

Kane, M. (2011). Validating score interpretations: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, *29*(1), 3-17. doi: 10.1177/0265532211417210

Koch, M.J. & DeLuca, C. (2012). Rethinking validation in complex high-stakes assessment contexts. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 99-116. doi: 10.1080/0969594X.2011.604023

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, *16*(2), 14-16. doi: 10.1111/j.1745-3992.1997.tb00587.x

Lissitz, R. W. &Samuelsen, K. (2007).A suggested change in terminology and emphasis regarding validity and education.*Educational Researcher*, *36*, 437-448.doi: 10.3102/0013189X07311286

Meckes, L. & Carrasco, R. (2010). Two decades of SIMCE: An overview of the National Assessment System in Chile. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 233- 248. doi: 10.1080/09695941003696214

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012- 1027. doi: 10.1037/0003-066X.35.11.1012

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: American Council on Education/Macmillan.

Moe, E. (2009). Introducing Large-scale ComputerisedAssessment Lessons. Learned and Future Challenges. In F. Scheuermann&J. Björnsson (Eds.), *The Transition to Computer-Based Assessment. New Approaches to Skills Assessment and Implications for Large-scale Testing* (pp. 51-56). Luxembourg: European Commission, Joint Research Centre Institute for the Protection and Security of the Citizen.

Moss, P. A. (2007). Reconstructing validity. *Educational Researcher*, *36*(8), 470-476. doi: 10.3102/0013189X07311608

Newton, P. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, *10*(1-2), 1-29. doi: 10.1080/15366367.2012.669666

Newton, P. (2013). Validity and the cultivation of valuable learning. In J. A. Baird, T. Hopfenbeck, P. Newton, G. Stobart& A. Steen-Utheim (Eds.), *State of the field review: Assessment and learning* (pp. 78-99). Oxford: OUCEA and Knowledge Centre for Education.

Nuñez, I. (2015). Educación chilena en la República: Promesas de universalismo y realidades de inequidad en su historia. Psicoperspectivas, 14(3), xx-xx. doi: 10.5027/PSICOPERSPECTIVAS-VOL14-ISSUE3-FULLTEXT-617

Ortiz, I. (2012). En torno a la validez del Sistema de Medición de la Calidad de la Educación en Chile [About the validity of the Quality Measuring System of Education in Chile]. *Estudiospedagógicos*, *38*(2), 355-373. doi: 10.4067/S0718-07052012000200022

Ministry of Education (1990).Ley OrgánicaConstitucional de Enseñanza (LOCE) N°18962, [Constitutional Organic Law for Education (LOCE) N° 18962]. Retrieved from

http://www.leychile.cl/Navegar?idNorma=30330&idVersion=1990-03-10

Sepúlveda, L. (2008). *El aporte del SIMCE a la discusión al interior de la escuela* [The contribution of SIMCE to internal discussions in schools].Santiago de Chile:Centro de Investigación y Desarrollo de la Educación, Universidad Alberto Hurtado.

Schiefelbein, E. (1998). Análisis del SIMCE y sugerencias para mejorar su impacto en la calidad [Analysis of SIMCE and suggestions to improve its impact on quality]. In S. Gómez (Ed.), *La realidad en cifras* (pp. 241-280). Santiago de Chile: FLACSO.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity.*Educational Measurement: Issues and Practice*, *16*(2), 5-8. doi: 10.1111/j.1745-3992.1997.tb00585.x

Sireci, S. (2007).On validity theory and test validation.*Educational Researcher*, *36*(8), 477-481. doi: http://edr.sagepub.com/content/36/8/477

Srivastava, P. & Hopwood, N. (2009). A practical iterative framework for qualitative data analysis.*International Journal of Qualitative Methods*, *8*(1), 76-84.

Stobart, G. (2009). Detemining validity in national curriculum assessments.*Educational Research*, *51*(2), 161-179.doi: 10.1080/00131880902891305

Taut, S., Cortés, F., Sebastian, C. &Preiss, D. (2009).Evaluating school and parent reports of the national student achievement testing system (SIMCE) in Chile: Access, comprehension, and use. *Evaluation and Program Planning*, *32*, 129–137.doi: 10.1016/j.evalprogplan.2008.10.004

Valenzuela, J.P., Bellei, C & De los Ríos, D. (2010). Segregación Escolar en Chile [School segregation in Chile].In S. Martinic&G. Elaqua (Eds.), ¿Fin de ciclo? *Cambios en la gobernanza del sistema educativo* (pp. 209-229). Santiago de Chile: UNESCO and Universidad Católica de Chile.

Wandall, J. (2009). National Tests in Denmark – CAT as a Pedagogic Tool. In F. Scheuermann&J. Björnsson (Eds.), *The Transition to Computer-Based Assessment. New Approaches to Skills Assessment and Implications for Large-scale Testing* (pp. 45-50). Luxembourg: European Commission, Joint Research Centre Institute for the Protection and Security of the Citizen.