Equity and Equivalence in Language Examinations

José Noijons jose.noijons@cito.nl
Cito - Institute for Educational Measurement, the Netherlands

Keywords: equity, equivalence, equating, language testing

Summary
Increasingly there is a concern that all students, young and old, should have an equal opportunity to enjoy an education that is most suitable to them to reach their aims in life. Equity is now at the centre of many education reforms: people should not be excluded from education, employment or health care because of traits that cannot be changed. In this paper we focus on equity and the use of the native language in examinations.
In countries where there is more than one national language there is often a requirement that national examinations are offered in all of these languages. For many subjects, such as Mathematics and the Sciences, this may be a minor issue. However, in the case of the First Language or Mother Tongue examinations this is a serious and complicated issue. Students in any of these national languages must have an equal opportunity to gain an equivalent score for the same ability. For this, examinations need to be made equivalent.
A related issue arises where examinations in foreign languages in a country will need to be equivalent. Many countries require student proficiency in at least one foreign language. Syllabi state that the same standards are required for any of these languages. But are examinations in these languages of equal difficulty and if not, how can this be achieved?
This paper reports on ways how Cito, the Dutch Institute for Educational Measurement has helped European countries in transition to achieve equity in examinations.

1. The need for Equity and Equivalence

Increasingly there is a concern that all students, young and old, should have an equal opportunity to enjoy an education that is most suitable to them to reach their aims in life and fulfil their ambitions. Thus in reforms of education systems issues relating to equity are paid much attention to. International initiatives, such as *Education for All* (EFA) have been underway to bring the benefits of education to every citizen in every society. In order to realize this aim, a broad coalition of national governments, civil society groups, and development agencies such as UNESCO and the World Bank committed to achieving six specific education goals:

1. Expand and improve comprehensive early childhood care and education, especially for the most vulnerable and disadvantaged children.
2. Ensure that by 2015 all children, particularly girls, those in difficult circumstances, and those belonging to ethnic minorities, have access to and complete, free, and compulsory primary education of good quality.
3. Ensure that the learning needs of all young people and adults are met through equitable access to appropriate learning and life-skills programs.
4. Achieve a 50 % improvement in adult literacy by 2015, especially for women, and equitable access to basic and continuing education for all adults.
5. Eliminate gender disparities in primary and secondary education by 2005, and achieve gender equality in education by 2015, with a focus on ensuring girls' full and equal access to and achievement in basic education of good quality.
6. Improve all aspects of the quality of education and ensure the excellence of all so that recognized and measurable learning outcomes are achieved by all, especially in literacy, numeracy and essential life skills.

In international surveys, such as the OECD's Programme for International Student Achievement (PISA) questions are asked to find out whether students are well prepared for future challenges; whether they can analyse, reason and communicate effectively and whether they have the capacity to continue learning throughout life. But PISA answers more questions. It looks at to what extent participating countries have a fair and inclusive education system that makes the advantages of education available to all. Education has expanded significantly in the past half-century, but hopes that this would automatically bring about a fairer society have been only partly realised. Women have made dramatic advances, but overall social mobility has not risen and in some places inequalities of income and wealth have increased. Increased migration poses new challenges for social cohesion in some countries while other countries face longstanding issues of integrating minorities. Fair and inclusive education for migrants and minorities is a key to these challenges. Equity in education enhances social cohesion and trust (Field, 2008).

For the Council of Europe, also, equal opportunities are at the heart of its language policies. Member states are faced with issues relating to national identity and other forms of identity, such as cultural and religious, as well as issues of social and political inclusion. In this perspective language education no longer seems quite so utilitarian, but more values-driven. *Plurilingualism* is therefore a policy goal to be developed as a value that is necessary for living together peacefully, inclusively and productively in our multicultural societies (Council of Europe, 2004).

In this paper we define *Equity* or *Equal Opportunity* as the circumstance that people are not excluded from education, employment or health care because of traits that cannot be changed; individuals are not placed at a disadvantage because they cannot choose their personal characteristics. Such characteristics may concern their gender, their age, their health, their special needs, their upbringing, their ethnicity, their colour of skin and/or the language(s) that they can or cannot speak. It is this last characteristic or trait that we will focus on in this paper: ethnicity and the use of the native language in examinations will be at the centre of this paper.

The second basic concept in this paper is *Equivalence*, the circumstance that two or more treatments, tests or procedures are essentially equivalent and any difference between them is of no practical or significant consequence. In any experiment, there will always be some difference in outcome when two treatments are applied or two different tests are administered. However, the question is not whether two tests lead to different results, the question is whether the outcomes differ enough to be scientifically relevant. And if they differ and are scientifically relevant, whether such differences are desired differences. In educational policy terms: whether differences in performance between, to give an example, two ethnic groups are desirable.

To study equity in education we may ask such questions as:

- What effect do the policies and the structure of education systems have on educational outcomes?
- Which school factors under the control of policy makers produce the best performance outcomes?

To answer such questions international surveys such as PISA look at how the structure of schooling (including the grouping of students, segregation of schools, management and financing, school resources, instructional climate) influence the quality and equity of educational outcomes and which school factors are associated with better quality and more equitable student performance.

2. Equity and Equivalence in Language Examinations

In many (European) countries there is more than one national language. And even if there is only one official (national) language this does not mean that no other languages are used in daily communication, as part of daily life, and as an integral part of people's identities. Countries have different ways to cope with this phenomenon in education. In some cases only the one national language is used in education and members of all linguistic minorities need to have mastered this national language to be able to successfully complete regular education. This may cause certain equity

challenges and some of the international organisations that were mentioned above have signalled these and have suggested ways to meet these challenges.

Other countries have given equal rights to both the majority language and to minority languages. As a consequence of this, in such cases (but not always) national examinations are offered in more than one national language. The challenge in this case will be that students with a background in any of these national languages must have an equal opportunity to gain an *equivalent* score for the same ability. In other words students who are equally "good" must have the same chance to the same grade, independent of their language background. To achieve this, national examinations in different national languages need to be made equivalent.

For many subjects, such as Mathematics and the Sciences, the existence of national examinations in different national languages may be a minor issue. It can be argued that the essence of these examinations is language-independent and that provided proper translation and back-translation procedures are in place, students with the same ability will have an equal chance to an equivalent score. Having said this, it is not without reason that in international surveys strict translation and back-translation procedures are being followed to avoid differences in performance due to non-relevant factors. Even then PISA and other international surveys have been challenged on translation issues (OECD, 2003). Typically, the translation process focuses on ensuring linguistic equivalence. However, establishment of linguistic equivalence through translation techniques is often not sufficient to guard against validity threats. In addition to linguistic equivalence, functional equivalence, cultural equivalence, and metric equivalence are factors that need to be considered when research methods are translated to other languages (Peña, 2007).

In the case of *Language* examinations in different national languages, it is a more serious and complicated issue to make sure that students with different language backgrounds who are equally good, have the same chance to the same grade. Language examinations often contain sections on sub-skills such as Grammar and Vocabulary; some examinations include sections on Literature. It is clear that to reach equivalence between such examinations is challenging. How can we make sure that Grammar questions in examinations in different languages are of comparable difficulty? How can we see to it that a task to interpret a poem in one language has a difficulty equivalent to such a task on a poem in another language? How can we make sure that there is cultural *equivalence* between such examinations when we accept cultural *differences* and indeed administer these examinations in different languages for the very reason that there *are* cultural differences?

A related issue arises where examinations in various *foreign* languages in a country will need to be equivalent. We can try and achieve equivalence by linking Foreign Language examinations to the Common European Framework of Reference for Languages (CEFR) developed by the Council of Europe (Council of Europe 2001). Thus a test at a particular CEFR-level in one language can be linked to a similar test at supposedly the same CEFR-level in another language through the medium of the CEFR. However, the linking process is rather complicated and the Council of Europe has published a Manual to help stakeholders in validating the linking process, in this case: validating equivalence between examinations (Council of Europe 2009).

Foreign Language examinations in various languages within one country may be based on an identical curriculum and/or syllabus and may be set at the same CEFR level. Yet this is no guarantee at all that these examinations are in fact at the level claimed. Very often links to the CEFR are not validated and it is not clear at all if these examinations are then of equal difficulty in terms of the CEFR. It is to be expected that from a *subjective*, student or teacher point of view the foreign language that is most familiar or most related is the least difficult, but that is beside the point here. It is also to be expected that both teachers and students will think that less frequently heard and taught languages are more difficult. They will tend to place examinations in such languages at higher levels than is warranted.

3 Methods of Linking

The procedures to achieve equivalence or comparability between two forms of a test, are called linking procedures. There are various methods of linking (Mislevy, 1992; Linn, 1993):

- *Equating:* the strongest type of linking (tests built to be the same in content and characteristics).
- *Calibration:* tests measure the same thing but perhaps with different accuracy or in a different way.
- *Projection:* test results predict scores on a different test that does not measure the same thing.
- *Statistical moderation:* relate test results statistically.
- *Social moderation:* judgments about comparability of performance level.

Equating is most frequently used when comparing the results of different forms of a single test that have been designed to be parallel. The US College Board equates different forms of the Scholastic Assessment Test (SAT) and treats the results as interchangeable. Equating is possible if test content, format, purpose, administration, item difficulty, and populations are equivalent. It is clear that in high-stakes national examinations that are important for a student's future, equation would seem to be the most suitable method of linking. In the case of parallel examinations in various national languages the content and item difficulty are the two most challenging elements that must be made equivalent. Even if ethnic populations are culturally different, in the above type of examinations such populations must be considered equivalent. The parallel examinations themselves must provide equal opportunities for all students.

Calibration can be done with tests that are constructed for different purposes, and use different content frameworks or test specifications. However, such tests will almost always violate the conditions required for equating. When scores from two different tests are put on the same scale, the results are said to be comparable, or calibrated. Most of the statistical methods used in equating can be used in calibration, but it is not expected that the results will be consistent across different populations. In the case of parallel examinations in various national languages the purpose of the tests may be identical and even the content framework may be comparable, but the test specifications will often be different and thus calibration may be needed in linking such tests.

Projection can be used to predict or *project* scores on one test from scores on another test without any expectation that exactly the same things are being measured. For all sorts of reasons it would be problematic to accept that a test that is not curriculum independent and is not language-specific can replace parallel examinations in various national languages. General language ability tests do exist, but for all sorts of validity reasons such tests would not generally be accepted as replacements of language tests in the national languages. However, such tests are sometimes used by the side of parallel examinations in various national languages to provide extra information.

Moderation is the weakest form of linking. It is used when the tests have different blueprints and are given to different, non-equivalent groups of examinees. Procedures that match distributions using scores are called *statistical* moderation links, while others that match distributions using subjective judgments are referred to as *social* moderation links. In either case, the resulting links are only valid for making some very general comparisons. We will see that we may have to revert to some of the above moderation techniques in the case of parallel examinations in various national languages as we are confronted with tests that have different blueprints.

As we wrote, in high-stakes national examinations *equating* would seem to be the most suitable method of linking parallel examinations in different *national* languages. Also, in the case of parallel examinations in *foreign* languages we need to equate. As we noted above, it would not be acceptable if an examination in Foreign Language A is more difficult than an examination in Foreign Language B when both examinations are based on similar/identical curricula and syllabi and both claim to be aimed at the same international proficiency level. In this context we thus have to equate between:

- different tests on the same subject administered to different populations (national languages);

- tests on different subjects based on the same curriculum standards to different populations (foreign languages).

An extra challenge - that we will not enter into here - is of course that we may need to equate between similar tests across years. This is necessary in such cases as nationwide tests at the end of secondary education that are used to qualify for university entrance. We must make sure that this year's candidates have an equal chance of getting an equivalent score when their proficiency is the same as or similar to lasts year's population.

4 Implementation of Equation

In this section we will briefly discuss methods of linking parallel national examinations in two national languages based on work that Cito, the Institute for Educational Measurement in the Netherlands, has carried out in some European countries in transition. In the context of World Bank Education Reform Projects in these countries Cito has contributed to founding Examination Centres and to introducing valid, transparent and equitable examination procedures. The examinations that received most attention were the so-called *Matura* examinations at the end of secondary school education. The challenge has been to make them acceptable as entrance examinations to institutes of tertiary education. This was to put an end to the practice of universities setting their own entrance examinations, which were not based on the curriculum, called for private tuition and were thus considered to be less equitable (Cito, 2008).

Many of the above-mentioned countries in transition have ethnic minorities that have been given the right to education in their native languages and consequently to examinations in their native languages. As part of the reform of the Matura examinations Cito was asked to help equate examinations across national languages and across years. This was a politically sensitive issue: ethnic populations of students may prove to have different abilities and thus the ability in equivalent subjects may be significantly different. As we pointed out above, this seemed less of a problem for Mathematics or the Sciences. If an ethnic group performed less well than another, this might be attributed to socio-economic factors, as international surveys have also been indicating. But for national examinations in the ethnic languages to show that one ethnic group is performing better in the native language than another group is more difficult to accept by stakeholders, even if such outcomes have also been found in international studies such as PISA and the Progress in International Reading Literacy Study (PIRLS).

Our specific aim has been to set equivalent *cut-scores (pass/fail scores)* on parallel national examinations in the different national languages. The way to do this has been to construct a *bridge* or *anchor* between two or more parallel examinations. A number or preconditions were formulated:

- There is an anchor of common items in examinations in language A and B.
- The anchor is representative of the whole examination.
- The anchor is sufficiently large (N of items).
- The anchor tests the same skill(s) in both languages.

So that students of equal ability in both populations will get equal scores on the anchor items, regardless of language.

According to the research literature (Mislevy 1992; Linn 1993) two tests can be equated using a third test as an anchor. This anchor test should have similar content to the original tests, although it is typically shorter than the two original tests. Often the anchor test is a separately timed section of the original tests. Sometimes, however, the items on the anchor test are interspersed with the items on the main tests. In our particular case, however, the anchor test was the longest section in each of the two parallel tests and it contained items that had been translated from language A into language B or the reverse. Care was taken that an equal number of original items was translated from A to B and the

reverse. We also made sure that the tests and tasks in this section were culturally neutral and referred to everyday phenomena that were equally present in the two languages.

Thus two parallel examinations contained a longer non-language-specific anchor section and a language specific section. The non-language-specific anchor section contained reading tasks on short and long continuous texts and on non-continuous texts (such as timetables, graphs, some advertisements). This section also contained a number of translated texts and tasks from the international (so-called *world*) literature. The language-specific sections contained questions on the ethnic Literature, and on Grammar and Vocabulary. From a content validity point of view this was acceptable. In the syllabus more emphasis is placed on the skill of Reading than on the sub skills of Grammar and Vocabulary. World literature is also a sizable part of the curriculum and the syllabus.

A separate score is then computed for the responses to those items as if they were a separate test. An assumption of the *equipercentile* equating methodology[1] that we have applied is that the linking function found in this manner is consistent across the various populations that are used in the equating. The research literature shows that this consistency is to be expected only when the tests being linked are very similar, which we believe they are.

A number of steps have been formulated for the above mentioned linking procedure.

Step 1
- Start from the examination with the largest number of candidates.
- Set a cut-score on the whole test, e.g. by using a modified Angoff method.
- Find the overall pass rate (%) for this examination.

Step 2
- Look at the distribution of scores for this examination but only for the common items.
- Find the cut-score on the common items corresponding to the same pass rate (%).

Step 3
- Transfer the cut-score to the distribution for the common items in the examination in the other language.
- Find the pass rate (%) on the common items corresponding to the same cut-score.

Step 4
- Transfer the pass rate to the distribution for the whole examination for the examination in the other language.
- Find the cut-score on the whole examination corresponding to the required pass rate.

A number of caveats are in place in the above procedure. Note that in the method described above the standards set in one language examination will determine the standards set in another. This may not be acceptable. Each language examination committee may set a provisional cut-score in advance. Then, the results from independent standard-setting are compared with those derived from the test equating procedure described above. Any discrepancies would need to be discussed in order to reach a consensus.

5 Conclusions

If we accept that students should have an equal opportunity to enjoy an education that is most suitable to them to fulfil their ambitions, we need to make sure that in high-stakes examinations students with

[1] Equipercentile equating adjusts the entire score distribution of one test to the entire score distribution of the other test.

different language backgrounds but with equal abilities have the same chance to be awarded the same grade. This will create challenges for (language) examinations in different languages which need to be resolved through equation of such examinations.

Before equation we must make sure that examinations are sufficiently similar to be able to ensure equivalence, with due respect for the uniqueness of each language and the culture in which it is used.

Standard-Setting procedures require good cooperation between examination officials representing relevant ethnic groups so that we can be sure that irrespective of which parallel examination has been administered candidates have equal opportunities to pass the examination. In this way we can contribute to equity in education.

References

Cito (2008). *Technical Assistance to the Examinations Sector of Macedonia.* Final Report. Arnhem: Cito.

Council of Europe (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment.* Cambridge: Cambridge University Press.

Council of Europe (2004). *Global approaches to plurilingual education; Summary Report.* Strasbourg: Council of Europe.

Council of Europe (2009). *A Manual for relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment*. Strasbourg: Council of Europe.

Field, Simon, Malgorzata Kuczera, Beatriz Pont (2008). *No More Failures: Ten Steps to Equity in Education.* Paris: OECD.

Linn, R.L. (1993). Linking results of distinct assessments. In: *Applied Measurement in Education*, 6, p83-102.

Mislevy, R.J. (1992). *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects.* Princeton, NJ: Educational Testing Service.

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

Peña, Elizabeth D. (2007). Lost in Translation: Methodological Considerations in Cross-Cultural Research. In: *Child Development,* Volume 78, Issue 4, p1255-1264.