The International Journal for Translation & Interpreting Research
trans-int.org

Development of reliable evaluation tools in legal interpreting: a test case

Heidi Salaets
KU Leuven, Antwerp campus
University of the Free State, Bloemfontein
heidi.salaets@kuleuven.be

Katalin Balogh
KU Leuven, Antwerp campus
katalin.balogh@kuleuven.be

DOI: 10.12807/ti.107203.2015.a08

Abstract: In recent decades, test design, assessment and evaluation procedures have received much attention and have focused on concepts such as quality, validity and reliability. Obviously this is also true for the highly complex testing of interpreters' skills, including legal interpreting. In this paper, we will first discuss the significant changes that have been made in the final examination procedure at the end of the LIT (Legal Interpreting and Translation) course at KULeuven, Antwerp campus, which have been complemented by an introductory workshop for the graders. It is important to mention that graders can be language experts as well as external legal experts (judges, prosecutors, police officers, lawyers, etc.) The comparison of the scores of candidates between 2008 and 2013 (a period in which different evaluation grids were used) shows a tendency towards more overall failures. In addition to this, an analysis of the graders' comments demonstrates that results are more consistent and that graders' comments mirror the results better. The new evaluation method clearly leaves less room for grader subjectivity, which presumably points to the fact that candidates are tested in a more transparent and reliable way. Follow-up research (in grader focus groups) and observations of the actual evaluation process will enable us to ensure that graders are comfortable with the new method and to check if they use it in a consistent way. Verifying whether the overall procedure actually produces better and more competent legal interpreters is a further important step needed to complete this research project.

Keywords: legal interpreting, evaluation procedure, reliability, validity, quality

1. Introduction

Before discussing the final examination in LI (Legal Interpreting) at KU Leuven, Antwerp campus, we would like to shed some light on the use of concepts without going into too much detail, however, or reinventing the wheel. We will also briefly explain the situation in Belgium and outline those elements which we will be unable to discuss in this contribution for reasons of space.

1.1 Certification through independent bodies

The final LI examination in combination with the other (written) examinations on legal knowledge, legal terminology, methodology and sources of law leads to certification, meaning that "certificates are usually awarded after completion of a course of study and demonstration of mastery of the knowledge or skills imparted in courses" (Mikkelson, 2013, p. 66). Furthermore, Mikkelson notes the following:

> Credentialing or certification, whereby mastery of the knowledge, skills, and abilities required to practice the profession is verified by an independent body, is inextricably linked to this formative education. (2013, p. 67).

Unfortunately, this is not yet the case in Belgium where no national register of sworn translators and interpreters exists and the title of legal interpreter is not protected. As a result, there are no standardized procedures allowing one to become a legal interpreter; each court has its own system for the recruitment of interpreters. The court of Antwerp alone has strict rules about the education, training, evaluation and certification of interpreters. This has come about as a result of the court's close collaboration with the LIT department of KU Leuven, Antwerp campus. Consequently, we are still light years away from well-designed specific certification tests like those in the United States (Feuerle, 2013) or other countries such as Australia, Austria, Canada, Sweden or the UK (Hlavac, 2013).

Nevertheless, the fact that an independent body exists in Belgium that certifies community interpreters, together with our experience as graders and contributors to the design of tests for community interpreters, were of great help in developing actual legal interpreting evaluation criteria. Vermeiren, Van Gucht and De Bontridder (2009) write about the certification of social interpreters (i.e. community interpreters) in the early years, while Roels (2013) explains how this certification process evolved, which led not only to a better and more valid test design, but also to training for graders and guidelines for more reliability in the testing procedure.

1.2 Assessment and evaluation

We wish to state clearly at this point that we will not comment in detail on the assessment procedure of the LIT course at KU Leuven, Antwerp campus. This would take in too much detail and would mean starting from screening during the admission procedure, moving on to feedback during class performances (formative assessment), and then to the final examination (summative assessment). Instead, only the LI evaluation at the end of the course will be discussed here because it rates the candidate's performance and achievements, while assessment procedures measure the performance and progression of an individual, giving feedback so that performance can improve.

This terminological difference between evaluation and assessment has been emphasized in many sources (see for example CIE, n.d.; PCrest, n.d.; ITLAL, n.d.) or kinds of assessment according to the time period(s) in which they take place). As expressed by the Institute for Teaching, Learning and Academic Leadership (ITLAL, n.d.):

> Assessment is the process of objectively understanding the state or condition of a thing, by observation and measurement. Assessment of teaching means taking a measure of its effectiveness. "Formative" assessment is measurement for the purpose of improving it. "Summative" assessment is what we normally call "evaluation." (http://www.itlal.org/index.php?q=node/93).

Although the citation refers specifically to teaching, that is secondary to what is fundamentally being assessed.

Formative assessments take place during the LI course and are designed to show candidates their strengths and weaknesses, giving them the necessary feedback and tools to remedy problems. Summative assessment on the other hand takes place at the end of the learning process and its purpose is to judge whether a candidate is ready to work in the profession. Since this assessment takes place during the final LI examination/role play, we are clearly talking about a final evaluation here. The fact that this evaluation is criterion-based and not norm-based will be discussed in paragraph 3.1.3.

1.3 Quality and evaluation

We wish to emphasize that the evaluation procedure used at KU Leuven, Antwerp campus, is not the only valuable one. As Pöchhacker states

> Quality is acknowledged as an essentially relative and multi-dimensional concept which can and must be approached with different evaluation methods from a variety of perspectives (2004, p. 153).

Taking different evaluation methods into account does not mean that evaluation should lapse into an almost personalized system of error counting or that measuring (legal) interpreting competences can be reduced to a mathematical calculation, without considering source-target correspondence, for example. This way, there is a risk that graders place their own interpretation on the classification schemes. If this occurs, evaluation reliability will be low (see 1.4). Furthermore, there are nearly as many error classification systems as there are empirical studies demanding an overall assessment of source-target correspondence (Pöchhacker, 2004, p. 143).

Of course, evaluation of an interpreting product and performance has also to take into account many features (Pöchhacker, 2004, pp. 137-158) such as discourse, source-target correspondence, effect, quality and role of the interpreter according to "the expectations held by participants in the interaction and in society at large" (Pöchhacker, 2004, p. 147). Therefore, a test design that allows for all these features is necessary to make a test valid (see 1.4).

1.4 Validity and reliability

An in-depth discussion of the concepts of validity and reliability in testing falls beyond the scope of this contribution and has already been undertaken by many authors. A very good summary that reflects the definitions and main issues on this topic can be found in Sawyer (2004, pp. 95-102). A brief overview can also be found in Salaets & Vermeerbergen (2011, pp. 164-166). A clear definition of the concept of validity reads as follows:

> To be considered *valid*, an assessment tool must test skills that are actually required to perform the task in question, and not test irrelevant skills; individuals who can do the job well should pass the test, and those who cannot do so should fail it […] (Mikkelson, 2013, p. 69)

Here, we will limit our discussion of validity to the above. Although serious efforts have been made to improve the test itself, i.e. the role play, we can only report that these efforts resulted in a better test design. In this newly designed role play, all competences that would-be legal interpreters should possess are tested in a well-balanced way. This contrasts with the old tests where role players had complete freedom in elaborating a broad scenario with only some key words. Dutch and foreign language proficiency as well as listening and speaking skills in both languages are tested. Interpreting techniques and skills are tested via the transfer itself and indirectly via note-taking techniques. Knowledge of law, legal terminology and the ethical code are, of course, tested through the legal context of the role play, whereas professional attitude is screened by means of a concrete ethical dilemma and interpreter behaviour during the complete role play (coping with stress, turn-taking, long(er) instances of speech, etc.). A plain definition of the concept of reliability is the following:

> […] a *reliable* assessment instrument is one that gives the same result for people of similar skill level regardless of who administers the test, who rates the test, when the test is given or what version of the test is applied (Roat, 2006, p. 9).

Angelelli illustrates even more clearly what is understood by the term:

> Reliability is not just about test score. Creating a reliable test and judging the reliability of an existing test involves looking at the ways in which the consequences of factors outside of what actually is being tested have been minimized to the greatest extent possible. (Angelelli, 2009, p. 17)

This is exactly what we wish to illustrate in this contribution, mainly through the analysis of the evaluation grids and guidelines, but also through a small scale study that shows how external factors, i.e. grader's freedom and manoeuvrability and thus subjective interpretation, can and must indeed be minimized to the greatest extent possible. Even when the test design is very well conceptualized and factors influencing reliability are reduced as much as possible, the next step to take - one which we were unable to do so far - is to develop and complete the testing cycle further whereby the "test itself is tested" (Leeson, 2013, pp. 157-161).

2. The evaluation process and procedure: a very brief history

2.1 Period 2000-2009

Every candidate took part in the entrance examination which consisted of a written and an oral Dutch test. There were no exemptions. The next part of the entrance examination tested the foreign language knowledge of the candidate interpreters and translators. Applicants who wanted to be legal interpreters participated only in the oral section of the language test, future translators in the written test. If candidates wanted to become both, they had to take both tests and prove that they had a thorough knowledge of the foreign language (both written and spoken). The requirements for the entrance examination were strict. In order to pass, candidates had to score at least 80%. After training in legal terminology, the Belgian legal system and law, students were trained in legal translation and legal interpretation. As the course was monolingual, it was not possible for students to improve their foreign language skills through the traditional way of teaching. Students improved their language level individually or together with other students (in small groups, if so desired), but without supervision or mentoring by a teacher. In short, there was no check on students' development and progress.

After training of some 160 hours, all candidates took an examination on legal subjects in Dutch. The translators had to translate part of a legal document, while the interpreters had to act as legal interpreters in a role play. The role play consisted of a simple story; the plot was not developed beforehand. The screeners received a brief description of the situation only, without any further instructions or structural guidelines.

From the start of the training program until 2009 the candidate legal interpreters and translators were assessed using the same evaluation grid. This meant no distinction was made between translators and interpreters and every screener used the same evaluation sheet, whereas it is beyond dispute that legal interpreters and translators need different skills and techniques. The evaluation sheet was divided into five main categories: The first was Dutch, with sub-categories that included correctness, vocabulary, language level of sentences and text. However, some other sub-categories applied only to interpreters, such as intonation, pronunciation and fluency. The next main category was foreign language usage, which contained the same sub-categories as the first category. Transfer was the third main category. This included completeness and correctness of the content, language skills, register and communication skills. Attitude was the fourth category. This comprised sub-categories such as stress tolerance, body language, accuracy, care/precision, reliability, objectivity, etc.

The final score was the last grouping in which the screener decided whether a candidate failed or passed.

The assessment sheet was cluttered. The evaluator himself had to decide whether a sub-category could be used and whether it worked for an interpreter or translator or both. The evaluator was not really involved in the process and was more like an external onlooker who had to assess the interpreting skills of the candidate by means of a chaotic and confusing document. Our research showed that the screeners did not fill in the sheet correctly, but through no fault of their own.

2.2 Period 2009-2011

In 2009 some initial changes were made to the final examination, while the entrance examination remained the same. A new step consisted of developing at least a different evaluation sheet solely for the legal interpreters. The screeners no longer had to stop and ask themselves which sub-categories were relevant for which profession. In addition, the categories mentioned earlier were now divided into additional sub-categories, i.e. omissions, additions, misunderstandings, ambiguity, ethics and specific terminology. The new evaluation sheet was still not perfect, but it opened the way for a much more structured and well-organized evaluation method. The evaluation sheet was subsequently further developed in 2010-2011.

One of the most important changes consisted of developing different evaluation sheets for the language grader and the legal expert. Another additional element in this new evaluation sheet was the greater spread in the scores. The score for each individual item was no longer based on a simple pass or fail. Both graders could choose the exact value for each category on a scale of 5-4-3-2-1 with 5 reflecting a very positive evaluation of the item and 1 a very negative one. The screener was more likely to distinguish good performances from poor performances and could also indicate whether the candidate's performance in one of the categories was excellent or merely mediocre.

3. The final examination from 2012 onwards

As explained in the introduction, we will not cover the entire assessment procedure but rather we will focus on the evaluative part, i.e. the final examination. While the test design had gone through a serious make-over, a following step still had to be taken to improve the grading of the interpreting examinations further. As mentioned above, the first change was introduced in 2010 by finally providing the two graders with a different grid, in line with their competences. The language grader was to assess the linguistic transfer in a detailed way as well as the attitude of the examinee and the correct use of the two languages tested (the foreign language and Dutch), whereas the legal expert was to assess the transfer in a more general way (observing misunderstandings and completeness), the attitude of the examinee, the competence in Dutch and specifically Dutch legal terminology.

When feedback was given during informal meetings, the legal experts in particular reported that the grids were difficult to work with as it was not entirely clear to them *how* they should be completed. Many of the terms were unclear to them. It was only with the help of and in agreement with the language grader that the legal expert's grid was completed, after which a final decision was taken.

That both graders work together to reach a common goal should naturally be encouraged. The fact that legal experts need some assistance in these matters is not at all surprising: they do not deal with grading and examinations every day. Unlike language graders, they are not trained in assessing students in such situations. While it is true that they often assess, they do this in a completely different manner and by means of different methods.

The above feedback made us think about how we could offer both the legal experts and the language graders a more workable tool that would guide them through the evaluation. As a result, we provided two solutions: we added guidelines to the grid and organized a workshop with all graders to explain to them how to work with the new grids and use these guidelines correctly.

3.1 The grids

The grids (see the appendix) require some further explanation which will be given in the following paragraphs.

3.1.1 The language grader's grids.

The most important task of the language grader is to check transfer, i.e. to check whether the content is fully conveyed in the target language. This takes place in the four phases of the role play, which correspond to four competences and allow the language grader to evaluate these four competences, namely the consecutive interpreting section (short and long), simultaneous whispering and sight translation. The grader assesses the accuracy and completeness of the transfer and also checks for major transfer errors (*contresens*).

Since it is not possible within the scope of this article to include all 10 pages of the guidelines, we will quote only the one guideline on accuracy that tries to answer the question of the grader: How accurate is accurate? In other words, what score do you give on the scale of 1 to 5? Scores 1 and 5 are clearly distinctive, as are 2 and 4, but it is more difficult to identify the difference between 5 and 4 on the scale, or 1 and 2. It is most important to know the difference between scores 3 and 2, because this signifies the huge distinction between a pass (3) and a fail (2). An example of part of the guidelines that accompany the grid follows.

> Think about changes/errors that have consequences for the information exchange. This is the case for example when:
>
> - one of the parties has to ask an additional question about something that was mentioned before but was translated accurately
> - there is wrong or inadequate use of general vocabulary, which makes information-gathering less accurate.
>
> **Example 1**: the use of hyponyms (a term that denotes a sub-category of a more general class)
>
> > "Did you hit him on the head with a chair? >> "Did you hit him on the head with a piece of furniture?"
>
> **Possible explanation**: The candidate cannot find the right word for "chair" and therefore uses "furniture" to complete the sentence. A chair is of course a piece of furniture, but this is not sufficiently accurate. This shift will lead to (at least) two possible mistakes: Misinformation if the suspect answers the question with "yes", while assuming that a "piece of furniture" can be anything small with which you can hit someone. The questioner automatically presumes that the piece of furniture is a chair because the answer is "yes". Or it can result in a surprising answer: "Yes, ehm, yes …with an ashtray".
>
> **Example 2**: there is wrong or inadequate use of legal terminology, which makes the information exchange less accurate or puts it at risk:
>
> > "Do you give permission to trace him? > "Do you give permission to arrest him?"
>
> **Example 3**: grammatical errors that radically change the content and put the information exchange at risk
>
> > Did your brother give you this? > "Did their brother give you this?"

Grading is as follows:
5: extremely accurate – no changes or distortions
4: maximum two changes – changes at word level without consequences for the information exchange and/or the conversation
3: two or three changes – less accurate, but things are put right during the conversation and there are no permanent consequences for the information exchange
2: four changes – not sufficiently accurate; the conversation/information exchange is at risk
1: more than four changes – not accurate; the conversation and information exchange have been clearly distorted.

Note-taking is also mentioned in this grid, but it is not assessed by the language screener as a separate skill and an end in itself. However, the interpreter's notes clearly have to be used as an effective tool to meet the other criteria of completeness, accuracy and correctness.

Finally, the grader is reminded that one fail, i.e. a 2 or 1 score, means a fail for the overall role play. If, by contrast, all competences receive scores of 3, 4 or 5, the candidate will pass for all these competences in the role play, which also means an overall "pass".

The two other language grader grids contain the language proficiency guidelines for Dutch and the foreign language. In Antwerp these criteria have been set at B2 level of the CEFR (the Common European Framework of Reference for Languages) for the simple reason that setting the standard at C1 or C2 level would result in a higher bar to pass and in having no legal interpreters at all. It is hard to pinpoint the exact difference between level B1 and B2 or B2 and C1. Probably nobody will readily have an exact description or definition of these levels in mind. At most, we can refer to level B2 as that of an 'independent user', but what does this mean exactly? In an attempt to standardize criteria, we outlined the characteristics of level B2 in a two-page summary. On the basis of these criteria, the grader has to decide whether the candidate meets the B2 requirements ('yes') or not ('no').

Firstly we cited the general definition of an independent B2 user which applies to the interpreter as follows:

- Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization.
- Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party.
- **YES / NO**

The main elements of the B2 CEFR level pinpointed in the respective definitions are:

Overall listening comprehension B2
- understanding a native speaker/interlocutor at B2 level

Overall spoken interaction B2
- information exchange at B2 level

At the bottom of the grid, we also give graders the opportunity to write down any remarks to justify their decision in the grid, i.e. simply 'yes' or 'no', reminding them of the requirements of the B2 level. Here again, one "no" leads to a fail for the items assessed: the foreign language and/or Dutch, both at B2 level.

3.1.2 The grids of the legal expert.

Now, after addressing the language experts' grids, we will take a closer look at the evaluation grid for the legal experts. The task of the legal experts (a policeman, judge, prosecutor or lawyer) is mainly to decide whether the candidate helps them to do their job properly and correctly. How can the legal experts in their turn contribute to the evaluation of the language graders? They can do so by supplementing the evaluation of the language graders with their own specific expertise and professional experience. Given the fact that legal experts might not be able to evaluate the language transfer as such, they are able to observe at a general level what information is conveyed in Dutch. They are able of course to compare this to the information they might expect to hear, bearing in mind that the examination is not a real life situation but a contrived situation where scenarios and the design of the test are played out.

In the consecutive interpreting section of the legal experts' grid, the graders can indicate how they assess the reliability of the candidate, including the candidate's attitude/communication skills, the use of the first person (in Dutch), impartiality (attitude and, if possible, content in Dutch) as well as the correct use of Dutch in general and of Dutch legal terminology in particular. It is easy to assess the use of the first person by keeping a tally of when the candidate uses indirect speech. It may happen on occasion that a candidate assumes the use of indirect speech of the witness ("tell the judge that I… and tell him also…"). For that reason, candidates are allowed to break this rule once or twice. If they wrongly use indirect speech instead of direct speech three or more times, this results in a (2) score (=fail). The legal expert also has to discuss with the language grader whether this happened in the foreign language as well.

It is, however, more difficult to grade the reliability of a candidate: How reliable is reliable? For this item, graders can also choose from scores ranging from 1 (lowest) to 5 (highest). Again, the most problematic score is the one each side of the dividing line between pass (3) and fail (2). For example, when the legal graders examine attitude and communication skills, they receive the following guidelines:

- The candidate displays a disturbing attitude that makes communication more difficult.
- The candidate is arrogant e.g. constantly interrupting the parties
- The candidate dominates the conversation e.g. taking the lead in turn-taking

or

- The candidate is shy: never dares to interrupt when necessary and longer instances of speech are possibly not fully rendered → discuss this further with the language grader
- The candidate is very insecure: constantly asks the speakers to repeat or frequently asks for additional information, clarification, etc.
- In general:
  The candidate is extremely nervous, is very intimidated by the situation.

The features described above can result in two possible attitudes: the interpreter is working very fast or very slowly → in both cases: discuss your opinion with the language grader.

Grading is as follows:

5: None of the above conduct was observed – very professional attitude/fluent communication

4: A maximum of two instances of the above conduct were observed – professional attitude

3: A maximum three instances of the above conduct were observed – sufficiently professional attitude provided that only minor attitude problems are evident and/or that these professional errors are corrected by the examinee

2: Three or four instances of the above conduct were observed – unsatisfactory professional attitude
1: More than four instances of the above conduct were observed – unprofessional attitude.

As can be observed, legal experts are explicitly asked to discuss their impressions with the language graders because it is felt that they can never be entirely sure that their impressions are correct – except in clear cases of attitude problems or when, for example, the interpreting takes 30 seconds while the original story lasted two minutes.

For the (simultaneous) whispering section and the sight translation, legal experts only are able to assess fluency, since these parts are translated into the foreign language. The concept of fluency is also clearly defined in the guidelines with the related scores (1 to 5). The second part of the legal experts' grid is the language proficiency grid, with guidelines for Dutch only - see 3.1.1. Finally, graders are also reminded that one fail (i.e. a 2 or 1 score) means a fail for the entire role play. If, by contrast, all competences are awarded scores of 3, 4 or 5, this means a pass for all the competences in the role play, as far as the legal experts are concerned.

3.1.3 The overall score document. The overall score document is meant to be used by both graders to write down their final decision: They indicate whether the candidate has a pass or a fail score for the foreign language (only the language grader assesses this) and for Dutch (both graders combine their observations). If they disagree on the scores, they will eventually have to take a decision and account for it, reaching joint agreement on the pass/fail since there can only be one outcome (only one option can be ticked here). Each grader then gives an overall score for the role play (pass or fail). This means that a candidate receives four scores in total and, as mentioned above, the candidate must have four pass scores to obtain an overall "pass".

Although this seems to be a fairly strict way of assessing candidates, we should bear in mind that in the case of this legal interpreting examination, we are giving new graduates immediate access to the labour market. From the moment they are sworn in, they can immediately start to work as professional interpreters in legal cases, which most likely will be more complex than the interpreting examination. It will certainly be less structured.

At this point, we would like to draw attention to the difference between norm-based and criterion-based evaluation. Academia and language graders are generally used to a norm-based evaluation which uses a ranking order they are familiar with. Students receive a score based on their degree of compliance with norm x, y, z. They can comply either a little more, much more or not at all. In most cases (language) graders then classify students on the basis of their competency level and give them a score of A, B or C or a 3, 4, 5 or a 12, 14 or 16 depending on the ranking system.

In this context, however, we are confronted with criterion-based evaluation: in this case to master the interpreting skills necessary to become a member of the profession. Compare it to the first time you fly a plane as a pilot: you can be very good at taking off, but at the same time you cannot allow yourself not to be proficient at keeping the plane in the air or landing. Interpreters cannot allow themselves to do a "fairly" good job, which may not be good enough to keep an innocent suspect out of jail.

4. Research Methodology

In our small research project we wanted to verify whether our evaluation method had an influence on the number of students who passed or failed, and whether we could rule out the subjectivity of the evaluator. The first question which could be addressed by using a quantitative research method was how the number of students who passed or failed had changed over the years, and more specifically from 2008 until 2013. The logical follow-up question was whether it was possible to identify any particular trends. If so, the obvious response would be to try and explain these trends. Moreover, we were interested to know if any tendencies could be attributed to the changes and adjustments to the evaluation sheet. While it remains difficult to demonstrate a clear-cut relationship between any of the observed trends and the changes in the evaluation sheet with the amount of data we will present, it is nevertheless possible to formulate some hypotheses.

A complementary qualitative research method, namely a more in-depth analysis of notes, observations and comments on the evaluation sheets, was needed to check if the evaluation criteria were sufficiently clear. The subsequent question was whether or not the aspects that underpinned positive or negative assessments were clearly described and defined. A further question was whether the graders expressed any subjective assessments and if so,

a) whether they corresponded to the competences of the grader, and
b) whether they were contradictory to the final score.

To find answers to the above, we studied the completed evaluation sheets of the examinations conducted between 2008 and 2013 (following the different evaluation procedures, as explained in paragraph 2). Scrutinizing every individual evaluation sheet was impossible; sometimes only one examination for a particular language, such as Malayalam (in 2009), had been conducted, and sometimes the number of examinations varied over the years, e.g. we had eight English exams in 2008, but not a single one in 2010. For this reason we chose two languages for which the number of examinations had remained stable over the years and which therefore allowed for a more balanced comparison of the results. We opted for two different categories of language: firstly, a school language - in our case French - and secondly, a language that is not taught in Belgian schools - Turkish. The former is one of the official languages of Belgium, along with Dutch (Flemish), German and Flemish Sign Language, and the latter is one of the most common and most frequently spoken languages in Belgium.

The number of examinations for these two languages appeared to be comparable over the years. We tried to ascertain whether the examination results were influenced by the four different methods (from 2008 to 2013), namely:

a) the old method (2007-2009);
b) the updated version with separate evaluation sheets for interpreters/translators (2009-2010);
c) the updated version with scales and separate grids for legal experts and language graders (2010-2011); and
d) the new method (used since 2012) with an evaluation guide for both graders (language and legal experts), a clear description of B2 level (CEFR) and training for the graders. For each type of sheet we examined in detail the main evaluation categories such as language skills (Dutch and foreign language), transfer, attitude and overall score.

We first conducted an analysis of these results from a quantitative angle and further interpreted the results on the basis of the qualitative observations.

The five quantitative outcomes revealed trends that came about over the years. The qualitative analysis on the other hand enabled us to interpret the written comments and remarks, and it also helped us to verify whether our hypotheses were correct.

5. Research results

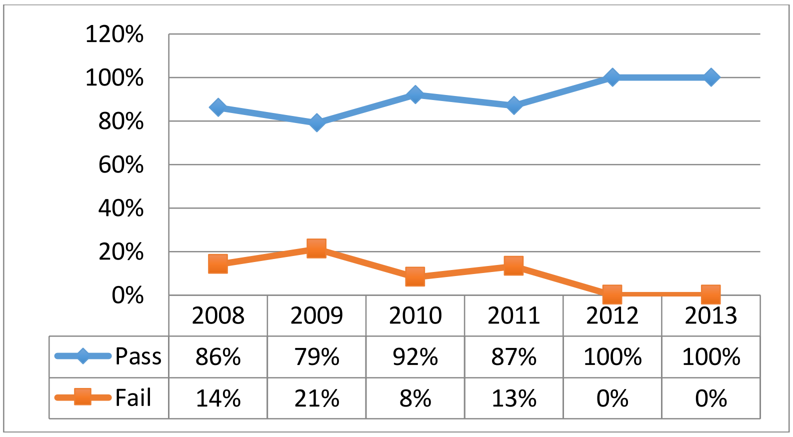
Below we outline the most important findings of the research. We have sub-divided this into a quantitative and qualitative analysis.

5.1 Quantitative analysis

As indicated above, we looked in detail at the five main evaluative categories in the interpreting examinations over the last six years (2008-2013). The outcomes always show the total score for both languages, French and Turkish.

5.1.1 Language Skills Dutch.

Passing the admission test for Dutch is one of the main requirements to follow the LIT course. This explains why the parameters (see figure 1) do not show huge differences concerning the Dutch language skills established at B2 level. Since 2012, however, no candidate has failed for Dutch and the graph shows a slightly upward trend. Not surprisingly, 2012 was also the year when all applicants had to pass both a written and an oral entrance exam, regardless of the course for which they were applying (interpreting or translating). Thus, from 2012 on, all future students had to pass the same written and oral entrance examination for Dutch.



	2008	2009	2010	2011	2012	2013
Pass	86%	79%	92%	87%	100%	100%
Fail	14%	21%	8%	13%	0%	0%

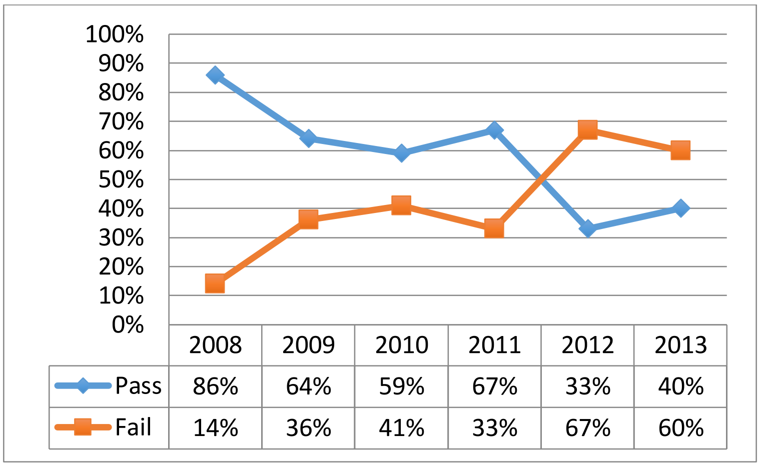
Figure 1. Dutch language skills

5.1.2 Foreign Language Skills (French and Turkish).

The most salient feature in figure 2 is the upward trend over the last two years (2012-2013). In the same period the number of passed candidates also increased. Since 2012, the number of failed candidates has been higher than the number of passed candidates. However, the huge discrepancy and variable gaps between the number of passed and failed exams have diminished.

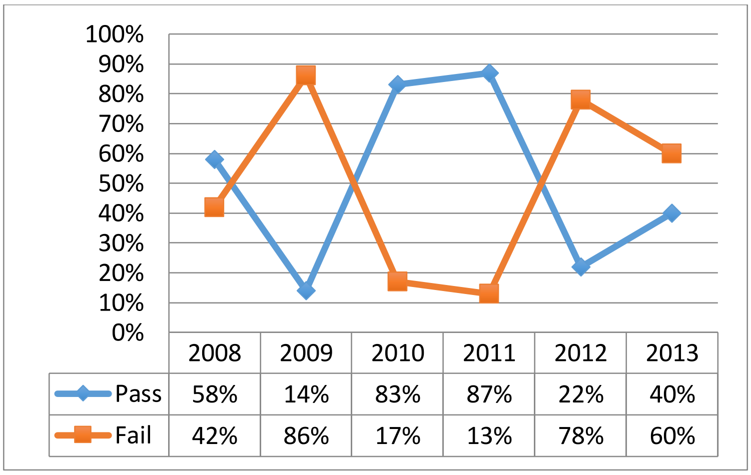
5.1.3 Transfer.

The results for the next main category, Transfer, paint a much more unbalanced picture. The line indicating the number of failed examinations shows huge ups and downs between 2008 and 2011, while the same is true for the passed exams. Only from 2012 on do both lines start to converge. Although the number of failed candidates is still higher than those who passed, the tendency is positive. A slightly upward (because only recent) positive trend can be observed.



	2008	2009	2010	2011	2012	2013
Pass	86%	64%	59%	67%	33%	40%
Fail	14%	36%	41%	33%	67%	60%

Figure 2. Foreign language skills (French and Turkish)

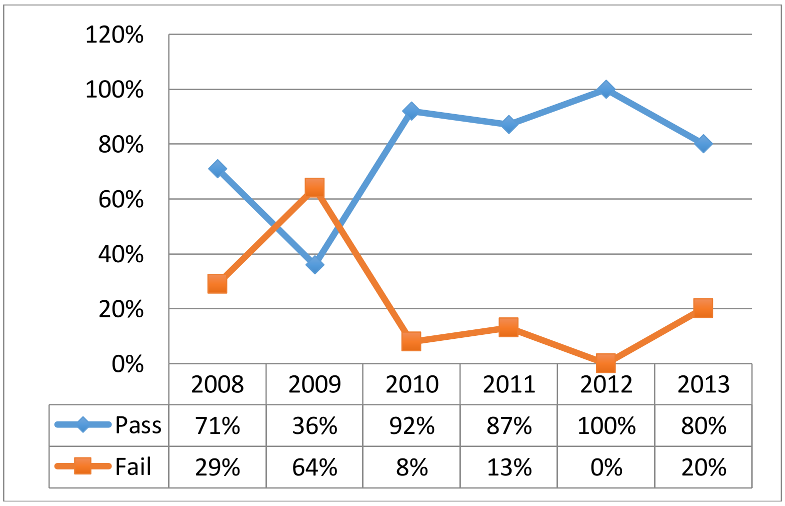


	2008	2009	2010	2011	2012	2013
Pass	58%	14%	83%	87%	22%	40%
Fail	42%	86%	17%	13%	78%	60%

Figure 3. Transfer

5.1.4 Attitude.

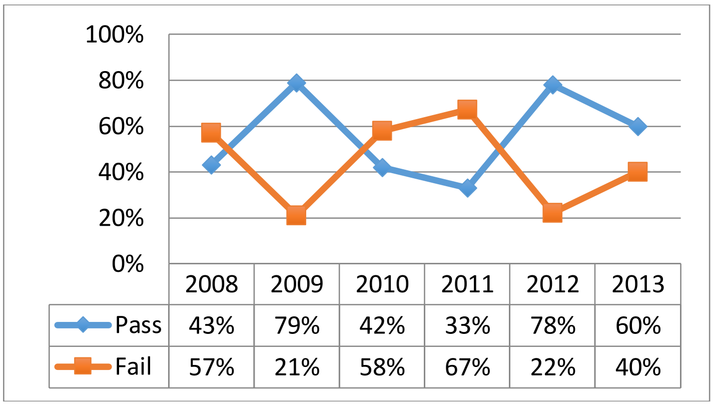
We can make virtually the same remarks for Attitude as for Transfer. Even though the gap between both lines is still extreme, we can see a positive and well-balanced trend from 2012 on. The extreme differences are tending to disappear.



	2008	2009	2010	2011	2012	2013
Pass	71%	36%	92%	87%	100%	80%
Fail	29%	64%	8%	13%	0%	20%

Figure 4. Attitude

5.1.5 Overall Score. The results of the overall score confirm the previous outcomes. The unbalanced and larger gaps have disappeared and from 2012 the lines have started to converge. This is perhaps the result of the cooperation between the legal grader and the language screener. We hope we can interpret this as an indication that they know more clearly what they need to focus on, and how and according to which criteria they have to decide whether a student passes or fails.



	2008	2009	2010	2011	2012	2013
Pass	43%	79%	42%	33%	78%	60%
Fail	57%	21%	58%	67%	22%	40%

Figure 5. Overall score

5.2 The Qualitative Analysis.

In the qualitative analysis we tried to answer three main questions by examining the evaluation grid as such. In the first place we wished to analyse the comments of the graders to see whether the evaluation was overly subjective and – to a certain extent - unreliable. The first qualitative question was: are the evaluation criteria sufficiently clear? The subsequent question was whether or not the aspects that underpin positive or negative assessments were clearly described and defined.

Results showed that these criteria were certainly not always clearly defined and that they were presented in groups in the early years (e.g. interpreting skills included fluency, transfer, etc.), embodying a holistic approach, which is a valid and generally accepted evaluation model. However, this changed in the new versions, as shown in the above sections. Guidelines to underpin positive (pass) or negative (fail) assessments were completely lacking in the early years. For the designers of the grid with scales and separate sheets for legal experts and language graders (in the year 2010), the exact meaning of the criteria and the scales seemed perfectly clear, but feedback from the graders showed that this was not the case.

Consequently, we checked whether graders wrote any comments to explain why they had taken specific decisions, since criteria were unclear or assessments were hardly underpinned. It became obvious that certain graders clearly explained and gave reasons for their decisions; others did not and some simply wrote “ok” three times (next to an assortment of undefined criteria). One wrote “not ok” and failed a student without further explanation. We can therefore conclude that the old evaluation method depended too much on the grader and that when criteria and evaluation scales were not clearly defined, it depended entirely on the attitude of the grader whether or not further clarifications were provided. This is not a reliable way of testing, because candidates can have good or bad luck with their grader. We also noticed, however, that the better the criteria are defined and the more that graders are familiar with how to use the scales (together with the guidelines), the fewer the comments that all graders add (there is no longer a discrepancy between graders who write more and those who write fewer comments). This hopefully means that the criteria and use of scales are clear for the graders, who therefore need to write fewer comments or explanations. In order to have a more detailed answer to this question, we need to organize focus group discussions and also individual interviews.

The second question we tried to answer was whether there were any subjective assessments expressed by the graders. The answer clearly is yes, but again the number of subjective assessments has been decreasing over the years. When we looked more closely at the comments, we noticed that some comments were contradictory to the final evaluation, but that this phenomenon only appeared in the early evaluation sheets (up to 2011). Some examples follow:

- a comment for the criterion transfer: *for longer instances of speech in consecutive, memory fails*
- a comment for the criterion attitude: *good attitude, aware of errors, corrects them if needed*
- a comment for foreign language skills: *hardly a pass mark*

Nevertheless, the overall score was a pass!

A candidate who failed in June 2010 because of bad results for whispered interpreting, nevertheless passed three months later despite the following comments:

- *no use of the first person in transfer*
- *interrupts the dialogue all the time*

One could legitimately ask how it is possible that anyone with such an unprofessional attitude can pass. With another grader, the same person would most likely/certainly fail because of the one important interpreting technique that was clearly not mastered (i.e. “NO use of the first person in transfer”) and an obvious attitude problem due to insecurity and/or arrogance (there are no further comments on why the dialogue was interrupted; nor do we know how frequent “all the time” is). This means that this evaluation method was not at all reliable.

The final aspect to be checked was whether the assessments and comments given by the graders corresponded to their own competences. We cannot answer this question for the period 2000-2010 since separate grids for the legal expert and the language grader were not used. In the period from 2010-2012, separate grids did exist but the evaluation criteria were not yet clearly separated according to the competences of the grader. This means that we regularly find a comment on an evaluation criterion that does not fall under the competence of the grader e.g. we find the legal expert commenting on foreign language knowledge or interpreting skills. In the last two years, graders have not only had separate grids, but also separate evaluation criteria: they write down the scores separately and only discuss them with their colleague when needed or when suggested by the guidelines. At the end of the examination, they discuss their scores together to establish the overall score.

6. Conclusion

Following this small-scale study, we can draw some conclusions that are far from general. They apply only to this sample and are closely related to the development of this particular curriculum, course and evaluation method.

Firstly, the redesigning of the role plays leaves less room for personal improvisation by the role players. Clearly designed scenarios and an introductory workshop on the importance of a valid and reliable evaluation procedure make role players – who are graders at the same time – more aware of the importance of a well-designed and rather rigid scenario to give equal chances to every examinee. They are also aware that it makes more valid testing possible, since the test assesses what it is supposed to evaluate, i.e. the examinees' language skills, general and legal knowledge, interpreting skills and professional attitude.

Secondly, the reliability of the final examination seems to be higher. If we look at the figures and scores for the different aspects (Dutch, foreign language, transfer, attitude and overall score), we notice a greater consistency for every item. This means that results have become more similar over the past two years, while they were clearly divergent in the years before, especially for such a rather vague criterion as attitude.

Furthermore, if we look at the qualitative analysis of the comments, we can see that the number of subjective or even contradictory assessments of the graders (when compared to the overall score) is decreasing and that graders, thanks to the separate grids, keep to the competences they are supposed to be assessing according to their expertise. All this makes us conclude that the graders' own attitude and subjective assessment is less likely to have an impact on the examination results because clearly defined criteria and extensive guidelines makes subjective interpretation of the grids less probable.

However, because of the limited amount of data, there is not yet sufficient evidence to show a strict and direct relationship between the changes made to the test/the evaluation system and the trends in examination results over the years. Furthermore, this research is still work in progress because the present study constitutes only a first step, by merely looking at the results and the graders' comments in the grids. Follow-up research should first ask the graders for their opinion on the grids, the guidelines and the role play scenarios (using focus group discussions and/or personal interviews). Video recordings of actual examinations and the evaluation process immediately afterwards should be made and examined in detail to see whether graders actually do what they say they do. Finally, the ultimate goal would be to check whether a better evaluation method actually results in legal interpreters who are more professional and competent than before, but this is of course a research project in its own right.

The most important conclusion that can be drawn from the current study is that graders are more certain of their ground and they can more easily account for their assessment on the one hand. Candidates, on the other hand, receive a substantiated overview of their strengths and weaknesses.

This makes us believe that, as Hilary Maxwell-Hyslop wrote in Building Mutual Trust (2011, p. 60), "(v)ery few people enjoy being assessed, but if candidates feel the process is transparent and fair, then they will, with luck, regard it as a necessary experience".

References

Angelelli, C. V., & Jacobson H. E. (Eds.). (2009). *Testing and assessment in translation and interpreting studies. A call for dialogue between research and practice*. Amsterdam-Philadelphia: John Benjamins.

Angelelli, C.V. (2009). Using a rubric to assess translation ability. Defining the construct. In C.V. Angelelli & H.E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies. A call for dialogue between research and practice* (pp. 13-47). Amsterdam-Philadelphia: John Benjamins.

CIE (Centre for Institutional Excellence, Purdue University). (n.d.). *Teaching Tips and Resources: Assessment and Evaluation*. Retrieved from http://www.purdue.edu/cie/teachingtips/assessment_evaluation/index.html.

CEFR (Common European Framework of References). (n.d.). http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Feuerle, L. (2013). Testing Interpreters: Developing, Administering, and Scoring Court Interpreter Certification Exams. *Translation and Interpreting. Special issue on certification*, *5*(1), 79-93. doi: 10.12807/ti.105201.2013.a04

Hlavac, J. (2013). A Cross-National Overview of Translator and Interpreter Certification Procedures. *Translation and Interpreting. Special issue on certification*, *5*(1), 32-65. doi: 10.12807/ti.105201.2013.a02.

ITLAL (Institute for Teaching, Learning and Academic Leadership, State University of New York). (n.d). *What is the difference between "assessment" and "evaluation?"* Retrieved from http://www.itlal.org/index.php?q=node/93.

Leeson, L. (2011). "Mark my words" The linguistic, social and political significance of the assessment of signed language interpreters. In B. Nicodemus & L. Swabey (Eds.), *Advances in interpreting research: inquiry in action* (pp. 153–176). Amsterdam-Philadelphia: John Benjamins.

Maxwell-Hyslop, H. (2011). The assessment of core competencies in legal interpreting and translation. In B. Townsley (Ed.), *Building Mutual Trust*. Retrieved from: http://www.buildingmutualtrust.eu/images/pdf/BMT-packaged.pdf

Mikkelson, H. (2013). Universities and Interpreter Certification. *Translation and Interpreting. Special issue on certification*, *5*(1), 66-78. doi: 10.12807/ti.105201.2013.a03

PCrest (Pacific Crest). (n.d.). *Learning module: Assessment*. Retrieved from http://www.pcrest2.com/LO/assessment/index.htm.

Pöchhacker, Franz (2004). *Introducing interpreting studies*. London-New York: Routledge.

Roat, C.E. (2006). *Certification of health care interpreters in the United States: A primer, a status report and considerations for national certification*. Los Angeles, California: The California Endowment. Retrieved from http://www.imiaweb.org/uploads/pages/195.pdf.

Roels, B. (2013). Certification of social interpreters in Flanders, Belgium: assessment and politics. In D. Tsagari & R. van Deemter (Eds.), *Assessment issues in language translation and interpreting* (pp. 179-197). Language testing and evaluation series (29). Frankfurt am Main: Peter Lang GmbH.

Salaets, H., & Vermeerbergen, M. (2011). Assessing students upon completion of community interpreting training: how to reshape theoretical concepts for practice? In C. Kainz, E. Prunč & R. Schögler (Eds.), *Modelling the field of community interpreting. Questions of methodology in research and training* (pp. 152-176). Wien-Berlin: LIT Verlag.

Sawyer, David B. (2004). *Fundamental aspects of interpreter education. Curriculum and assessment.* Amsterdam-Philadelphia: John Benjamins.

Vermeiren, H., Van Gucht, J., & De Bontridder, L. (2009). Standards as critical success factors in assessment: certifying social interpreters in Y. In C.V. Angelelli & H.E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies. A call for dialogue between research and practice* (pp. 297- 329). Amsterdam-Philadelphia: John Benjamins.

Appendix 1
Role play Lit-Language Grader – Evaluation Grid

Transfer Dutch-FL-Dutch		**5 very good**	**4 good**	**3 sufficient**	**2 insufficient**	**1 completely insufficient**
SHORT CONSEC	completeness					
	accuracy					
	capital transfer error	*absent*	*absent*	***one***	*more*	*frequent*
Transfer Dutch-FL-Dutch		**5 very good**	**4 good**	**3 sufficient**	**2 insufficient**	**1 completely insufficient**
LONG CONSEC	completeness					
	accuracy					
	note taking technique					
	capital transfer error	*absent*	*absent*	***one***	*more*	*frequent*
Transfer Dutch-FL-Dutch		**5 very good**	**4 good**	**3 sufficient**	**2 insufficient**	**1 completely insufficient**
SIGHT TRANSLATION	completeness					
	accuracy					
	capital transfer error	*absent*	*absent*	***one***	*more*	*frequent*
Transfer Dutch-FL-Dutch		**5 very good**	**4 good**	**3 sufficient**	**2 insufficient**	**1 completely insufficient**
SIM WHISPERING	completeness					
	accuracy					
	capital transfer error	*absent*	*absent*	***one***	*more*	*frequent*
REMARKS	(write down below this line)					

Appendix 2
Role play Lit-Legal Expert/Grader – Evaluation Grid

Professionalism		5 very good	4 good	3 sufficient	2 insufficient	1 completely insufficient
SHORT CONSEC	reliability					
	attitude/ communication					
	first person (“I”)					
	impartiality					
	legal terminology/ knowledge					
Professionalism		5 very good	4 good	3 sufficient	2 insufficient	1 completely insufficient
LONG CONSEC	reliability					
	attitude/ communication					
	first person (“I”)					
	impartiality					
	legal terminology/ knowledge					
Professionalism		5 very good	4 good	3 sufficient	2 insufficient	1 completely insufficient
SIGHT TRANSLATION	fluency					
SIM WHISPERING	fluency					
REMARKS	(write below this line)					