The parrot-exam: How valid is it?

Qualitative research into the validity of the TGN in the Netherlands using Bachman & Palmer’s (1996) model for the evaluation of usefulness.

Anne H.J.M. Abeling
S2354209

MA in Applied Linguistics
Faculty of Liberal Arts
University of Groningen

Supervisor: Dr. H.I. Hacquebord
Second reader: Dr. M.H. Verspoor

July 10, 2013

Word Count: 16,459
Abstract

This paper scrutinises the validity of the Test of Spoken Dutch (Toets Gesproken Nederlands, TGN), which is incorporated in the Civic Integration Exam, by applying the model for the evaluation of usefulness (Bachman & Palmer, 1996). The model defines validity by six test qualities (reliability, construct validity, practicality, authenticity, interactiveness, and impact) which should all be maximised in tests where the stakes of passing or failing are high. The washback of the TGN on teaching and learning is looked into as well by means of interviews with test takers and teachers in the field. The results show that the validity of the TGN is low on construct validity, authenticity, interactiveness and impact. Moreover, the test creates negative washback. The results, which were combined with reviews from a panel of experts, suggest that the TGN is not valid for its use in the high-stakes Civic Integration Exam as the only measurement of spoken language ability.
Acknowledgements

This thesis could not have been written without the help of all the test takers and teachers who let me interview them. These interviews are a core part of this thesis. Therefore, I would like to thank all teachers and test takers for their cooperation.

In addition, the panel of experts, who provided valuable feedback on parts of this thesis and who willingly shared their precious time, deserve special gratitude. The quality of this thesis would not be what it is today were it not for them.

A special word of gratitude to Eefje Cadée, Karen Heij, Mirna Pit, and many others from Bureau ICE, who introduced me to the field of testing, but also managed to raise my interest in testing during my internship. Without them I would probably have never come up with the current subject of research.

I would like to thank my family, friends, and fellow students for their support, understanding, and contributions to this thesis. It would not have been the same without you.

Last but certainly not least, I would like to thank my supervisor, Hilde Hacquebord. Thank you for your feedback in my search for good research and all other stages these past months.
ABBREVIATIONS

CGN – Corpus Gesproken Nederlands (Dutch acronym for the Corpus of Spoken Dutch). This corpus contains transcribed data of about 900 hours of spoken contemporary Dutch (from both the Netherlands and Flanders). The corpus is made up of almost nine million words.

TGN – Toets Gesproken Nederlands (Dutch acronym for Spoken Test of Dutch). This test is used in the Dutch Civic Integration exam (inburgeringsexamen) which tests citizens to-be on the A2 level.

STEXI – Staatsexamen I (Dutch acronym for State Exam Program I) the higher levelled B1 exam in the naturalisation program). This exam needs to be completed if the citizen to-be wants to proceed with vocational education.
# Content

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>3</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>4</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>5</td>
</tr>
<tr>
<td>Introduction</td>
<td>7</td>
</tr>
<tr>
<td>Theoretical Framework</td>
<td>8</td>
</tr>
<tr>
<td>Validity</td>
<td>9</td>
</tr>
<tr>
<td>Evaluation of usefulness</td>
<td>9</td>
</tr>
<tr>
<td>Washback</td>
<td>12</td>
</tr>
<tr>
<td>Validity in computerised tests</td>
<td>15</td>
</tr>
<tr>
<td>Statement of purpose</td>
<td>17</td>
</tr>
<tr>
<td>Methodology</td>
<td>18</td>
</tr>
<tr>
<td>Subjects</td>
<td>18</td>
</tr>
<tr>
<td>Materials</td>
<td>19</td>
</tr>
<tr>
<td>Procedures</td>
<td>20</td>
</tr>
<tr>
<td>Results</td>
<td>22</td>
</tr>
<tr>
<td>Analysis TGN</td>
<td>22</td>
</tr>
<tr>
<td>Interviews</td>
<td>27</td>
</tr>
<tr>
<td>Evaluation of usefulness</td>
<td>33</td>
</tr>
<tr>
<td>Washback</td>
<td>40</td>
</tr>
<tr>
<td>Panel-review</td>
<td>41</td>
</tr>
<tr>
<td>Discussion</td>
<td>42</td>
</tr>
<tr>
<td>Evaluation of usefulness</td>
<td>42</td>
</tr>
<tr>
<td>Washback</td>
<td>45</td>
</tr>
<tr>
<td>Panel-review</td>
<td>46</td>
</tr>
<tr>
<td>Conclusion</td>
<td>46</td>
</tr>
<tr>
<td>References</td>
<td>49</td>
</tr>
<tr>
<td>Appendices</td>
<td>54</td>
</tr>
<tr>
<td>Practice test TGN</td>
<td>55</td>
</tr>
<tr>
<td>Interview questions</td>
<td>58</td>
</tr>
<tr>
<td>Bachman &amp; Palmer’s model</td>
<td>59</td>
</tr>
</tbody>
</table>
Introduction

The Netherlands is one of the only countries in the world which requires applicants for immigration into the Netherlands to pass three stages before becoming ‘fully Dutch’: admission to the country, integration into the country and becoming naturalised as a citizen (Extra & Spotti, 2009). Since the Law on Integration passed in 2006, the second stage, in which a series of exams need to be passed, has received a considerable amount of negative criticism. One of the reasons is that immigrants have to arrange and finance the exams themselves. Moreover, the exams need to be passed within three years in order for the immigrant to avoid being sent back to their country of origin. This makes the test high-stakes as a lot depends on passing the exams. Although there is considerable general criticism on the Law, most criticism has been, and still is, on one exam in particular: The Toets Gesproken Nederlands (Test of Spoken Dutch, hereafter named TGN). The TGN is an automatically scored exam, administered via telephone, testing listening and speaking abilities on the A2 level of the Common European Framework of Reference (Council of Europe, 2001). The exam consists of four parts: repeating sentences, giving short answers to questions, giving antonyms and retelling stories. As most items consist of repeating what is heard, the exam is commonly referred to as the ‘parrot-exam’.

The TGN has received abundant negative criticism from different directions. In January 2013 an episode of the TV-program called Kassa¹ was devoted to this test. Interviewees on the program (test takers, teachers, researchers and a political party) claimed that the TGN is not a valid test because people who speak Dutch at the A2 level (as was stated in the TV-show by speech-therapists and teachers) did not pass the test. Therefore, they claim, there must be something wrong with the test. Later that year (in March) a newsprogram (RTL nieuws) stated that people who do not speak Dutch at the A2 level can pass the exam. Both programs thus claim the test is not valid, though for different reasons. Moreover, discussion on LinkedIn and articles in the Dutch magazine Les² address more or less similar issues with the TGN in the field.

Altogether, it is clear that there are validity issues in the field concerning the high-stakes TGN. These issues are the starting-point of this thesis which will look deeper into the validity of the TGN by using a model proposed by Bachman & Palmer in 1996. This model is defined by the authors as an evaluation of test usefulness. It focuses on the usefulness of the interpretation of the score and is defined by the following types of validity: reliability,

---

¹ TV-program in which complaints of consumers are solved by mediating between two parties (kassa.vara.nl)
² Dutch magazine for teachers of Dutch as a Second Language
construct validity, practicality, authenticity, interactivity, and impact. It is important to note that Bachman & Palmer (1996) have their own way of seeing validity (which will be discussed in the next chapter in more detail). Moreover, besides a discussion and evaluation of these types of validity this thesis also includes stakeholders’ opinions about the TGN. Teachers and test takers are interviewed for the purposes of getting insight into both the impact of the test on them and the effect of the test on teaching and learning. This thesis will therefore look into the validity of the TGN by using Bachman & Palmer’s (1996) model for the evaluation of usefulness. It will also provide insight into the effect of the TGN on teaching and learning (a phenomenon known as ‘washback’).

Underneath, an chapter 2 will give an overview of the research into validity and washback is given, which serves as the basis to get insight into the possible validity of the TGN. This section will also describe the Bachman & Palmer (1996) model for the evaluation of usefulness. The research done on validity in both automated and non-automated tests is described as well. Chapter 3 describes what methods were used to research the validity and washback of the TGN followed by the methodology used in the panel-review. Chapter 4 describes the evaluation of usefulness for the TGN as well as the results from the panel-review. Chapter 5 then discusses these results, which are used to answer the research question in Chapter 6.

Theoretical Framework

The testing of language abilities has been present in society for a very long time. Shibboleth-tests were used in ancient history to identify people from a certain class by means of the word Shibboleth which was pronounced differently by two peoples in the Hebrew Bible (McNamara & Roever, 2006). More modern usage of such tests happened in World War II, when people were asked to pronounce the Dutch city name ‘Scheveningen’ to identify who was Dutch and who was German. The answers to these language tests could lead to serious consequences: being arrested or even murdered.

Even though the relation between language abilities and how these abilities are tested has changed and evolved considerably (see Fulcher, 2010), it is important to realise that (although debatably not as severe as in Shibboleth-tests) test-results can still have serious consequences. Tests can either open or close doors. Failing a test might mean that you are not allowed to go to university or that you are sent back to your home-country. The stakes or consequences of these tests are high: a lot depends on passing the test. Tests like these are therefore called high-stakes tests and are usually instated by external organisations such as the government (Fulcher, 2010). The test which is of interest in this
thesis (the TGN) is a test with high-stakes. If test takers do not pass the test within three years, they will not be allowed to immigrate into the Netherlands. High-stakes tests usually cause high levels of anxiety and stress, which in turn impact performance (Embse & Hasson, 2012), and reduces hours of sleep (Fulcher, 2010).

Thought has to be put into the development and implementation of the test because of the high stakes a test might have. More precisely, the test or test scores should be valid. The meaning of the terms ‘valid’ and ‘validity’ is explained next.

Validity

A valid test, in its original definition, is a test which measures what it intends to measure (Valette, 1967). This definition has been further defined and refined by many researchers (for a full overview see Chapelle, 2012), which led to a diverse set of different types of validity: content validity, empirical validity, face validity and predictive validity (among others). These types of validity were united in a single definition by Messick (1989), who defined validity as ‘an overall evaluative judgement of the degree to which evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores’ (emphasis in original, Messick, 1989: 14). Messick’s insights were different from previous definitions in that he ascribed validity not as a property of a test but rather as the interpretations made from test-scores. Moreover, he united all different validities under one umbrella term, ‘validity’, with construct validity as the central one. Furthermore, he brought the old technical concept of validity to a more social concept which ‘encompasses the relevance and utility, value implications and social consequences of testing’ (Chapelle, 2012: 24). Finally, Messick saw the validation of an interpretation of a test or its use as a perennial process.

With this new definition of validity a framework was needed which could be used to assess how valid an interpretation of a test was. Bachman & Palmer (1996) opted for such a model in which they took Messick’s definition a step further in that it could be applied in test development, but also in reviews of current tests. The model is explained below.

Evaluation of usefulness

Bachman & Palmer (1996) relate their testing practice to current ideas on communicative language teaching and believes that this link should always be present. That is, they believe that the best way of language testing is in line with communicative language teaching. Therefore, authentic tests are their preference. It is important to realise that their complete theory on sound language testing is based on the degree of correspondence between language test performance and actual language use.
The model for the evaluation of test usefulness includes six types of validity (or test-qualities) which need to be maximised. The qualities are: reliability, construct validity, practically, authenticity, interactiveness, and impact. Bachman & Palmer set up 41 questions which are connected to the six qualities. All questions should sufficiently be answered in their model. The questions and the model can be found in the appendix. This model and corresponding questions are used in evaluating the usefulness of the TGN which is reported later in this thesis.

According to Bachman & Palmer (1996) the six qualities should be balanced. This balance varies between tests. It should be noted that complete satisfaction on all six qualities cannot be achieved, although especially in high-stakes tests all qualities should be maximized. The model can be used to answer the question: ‘How useful is this particular test for its intended purpose(s)?’ (1996:17). This also means that a test can be valid for one use, but not for the other. Validation is an on-going, perennial process which lies on a continuum. That is, the interpretations are never absolutely valid.

Moreover, the six qualities should not be evaluated on their own but be combined. For example, Bachman & Palmer (1996) claim that interactiveness, authenticity, and construct validity are dependent upon how the language ability is construed in the test. This means that if the construct or language ability is ill-defined, all three qualities are affected. Furthermore, the impact a test may have is dependent upon the test characteristics and its authenticity. That is, negative impact on instruction can be minimized by increasing the authenticity of the test. In fact, Bachman & Palmer (1996) state that the test-developer should focus on increasing the authenticity to decrease negative impact. All six qualities are explained next.

**Reliability.** This quality of validity essentially is the ‘consistency of measurement’ (Bachman & Palmer, 1996:19). When person A makes the same test twice in a short amount of time the scores should be more or less equal. This is an essential condition for any test: we need reliable information about the language abilities of a person in order to be able to make decisions. Reliability can be linked to previous types of validity as criterion-oriented, predictive and concurrent validity.

**Construct validity.** This quality refers to the test’s basis: the construct. What measurement do you use to get insight into the language ability you want to test and is that justifiable? ‘That is, we need to demonstrate, or justify, the validity of the interpretations we make of test scores, and not simply assert or argue that they are valid’ (Bachman & Palmer, 1996:21). Construct validity refers to both the construct definition and the tasks being used. It makes the link between test use and actual language use. That is, once the construct is defined, authenticity and interactiveness can be looked into as the construct is likely to influence these two. Thus, construct validity is a broad test-quality and can be justified with
evidence from the test’s content, relevance and coverage among others. More evidence is needed when the test is high-stakes. Although before this test quality could previously be found as well, construct validity can also be related to previous definitions of content validity and criterion-related validity.

It should be noted that both validity and reliability are necessary conditions for usefulness. A test can give highly reliable scores but might not be valid in its construct and vice versa. In both instances the test is not very useful.

**Practicality.** This type of validity does not necessarily refer to the use of the test but rather refers to its implementation, resources, design, actual use etc. It is important to note that a test is either or not. In the case of the TGN this means that the test fits the ideas of the Ministry which provided all resources. Practicality cannot be linked to any previous type of validity.

**Interactiveness.** This quality has to do with the interaction between test taker’s individual characteristics and the test’s characteristics. It refers to whether the test takers have to use their language ability, knowledge of the world and affective schemata when completing the test. This type is expressed on a continuum: it is not an either or type of validity. An example of a test which is not very interactive would be a test which tests whether a person can read graphs: no language is needed for this. Interactiveness can more or less be related to content and ecological validity.

**Authenticity.** Authenticity can be defined as the amount of correspondence between the test task and a real-life task, which can be related to previous definitions of content, criterion-related, and ecological validity. This quality is not generally seen as an essential quality of language tests. However, Bachman & Palmer (1996) consider the quality to be important as it combines the test situation with real-life settings. Tests are generally used to get insight into language abilities and to generalize these to what a person might be capable of in actual language use. The test is likely to give a better indication when it is highly authentic. Also, the degree of correspondence between test-use and actual language-use is potentially beneficial for test takers’ perceptions of the test.

Another important reason for looking into the authenticity of the test is because it relates to the motivation or impact the test may have on the test takers. Test takers might not see the use of the test and thus are not very willing to take the test. If the test is relevant to the skill (or TLU), a more positive attitude towards the test may be found. This relates to another type of validity: impact.
Impact. The impact that a test may have on different stake-holders (i.e. individuals, teachers and entire (educational) systems) can be rather large. Preparing for the test can be strenuous (or very easy). Also, the decision made when the test is administered can have a large impact as well. Especially when the test is high-stakes. Impact can occur on three aspects for test takers: in the preparation for and administration of the test, the feedback received and the decisions made. Obviously, these steps have an impact on the teacher and perhaps even larger systems. Impact can be related to consequential validity but face validity as well.

An important side-effect of the impact a test might have can be found in the teaching and learning in preparation of the test. The side-effect is washback and will be discussed in more detail in the next section as Bachman & Palmer (1996) do touch upon this phenomenon but fail to incorporate it sufficiently in their model. The next section explains its importance in the evaluation of test validity.

Washback

The introduction showed considerable negative criticism from different fields on the TGN. On the one hand it was stated that people who do seem to speak Dutch at the A2 level did not pass the test and vice versa. Moreover, the discussions on LinkedIn and Les showed that test takers, teachers, and researchers do not agree with this test. Therefore, it is interesting to see what the effect of a test is on teaching and learning.

Washback generally occurs when the test is high-stakes. Washback is the influence of the test on learning and teaching and happens when students and teachers ‘do things they would not necessarily otherwise do’ because of the test’ (Alderson & Wall, 1993:117). That is, what is in the test is more likely to be taught at the expense of other parts of language ability (Wall, 2012). Washback can either be positive or negative (Bailey, 1996). Positive washback may be found when teachers use tests to motivate their students to work hard in class (even though the test might not be very valid) (Alderson & Wall, 1993). Messick adds that ‘for optimal positive washback there should be little, if any difference between activities involved in learning the language and activities involved in preparing for the test’ (1996:241). Although theoretically positive washback is possible, hardly any research found clear positive washback effects (Cheng, 2010). Most research found negative or mixed results.

Negative washback may be found in anxiety in the learner or concern from the teacher. In the former case the negative washback may lead to lowered performances. In the latter case, teachers may want their pupils to pass the test and thus start teaching to the test. Thus, tests can be used to promote the learning of certain aspects of language or to impede
it. For example, a test focussing solely on the writing of the past simple will likely encourage teachers to focus on the past simple in class. On the other hand: if the test includes all tenses of the English grammar, classes will be tailored to this as well. Ideally then, teaching and testing should involve the same activities, i.e. using authentic materials in the test so beneficial washback is created (Messick, 1996).

Relatively little research has looked into washback, even though most research agrees on the effect that testing can have on teaching and learning. Some researchers consider washback a type of validity on its own (Morrow, 1986), although Morrow did not know how to tackle this type of validity nor could he prove its existence. Messick (1989; 1996) considers washback part of a test’s consequential validity. He does note that it should be evidentially shown that the washback is in fact a consequence of the test. If this is the case, it should be weighed in the evaluation of validity or usefulness. Anderson and Wall’s (1993) review of research into washback was the first attempt to bring the phenomenon in perspective. They report one study in which hardly any negative washback was found (Kellaghan et al., 1982), which was most likely due to it not using an actual (high-stakes) test. Another study did show results (Smith, 1991) by using interviews and classroom observations. Thus, the research up till then presented different results, which is likely due to their experimental set-up.

More recently, Cheng, Watanabe & Curtis (2004) published a book on washback in language testing which describes washback as a complex phenomenon. Watanabe (2004) advocates for qualitative research into the phenomenon. After identifying the problems with the test (for example through ethnography), interviews may be undertaken or observations may be done. Preferably, a combination of both is used. Burrows (2004) used a combination of surveys, interviews and observations in research into washback in Australia’s Adult Migrant English Program. She found a link between washback and curriculum innovation but notes that teachers have a choice in washback – they may implement the test into their teaching but this is not necessary.

Stecher, Chun & Barron (2004) used surveys to get insight into the teaching of writing after assessments had been instated by the government in Washington. Teachers in the system reported changes in time-allocation on writing in class. They found strong evidence that non-tested abilities were not taught. Thus testing seems to drive teaching in this context. Hayes and Read (2004) looked into the washback of university-entrance tests in New Zealand by means of questionnaires, interviews and a classroom study. Their study found clear washback effects: students and teachers focussed on what was required in test tasks rather than increasing their overall proficiency. Another school, however, did focus on overall
proficiency. The main difference between the two was time allocation. The first school only had four weeks to prepare for the exam, the second eight months. This would indicate that a shorter preparation time influences negative washback. Finally, Ferman (2004) looked into washback on an EFL oral test in Israel through questionnaires and found positive as well as negative washback, in the promotion of learning and the narrowing of what is taught respectively. Thus, Stecher et al. (2004), Hayes & Read (2004) and Ferman (2004) found clear washback effects.

However, this research did not include stakeholders’ perspectives on test experience and perceptions of test validity (Cheng & DeLuca, 2011), although both Messick (1989) and Bachman & Palmer (1996) advocate for the inclusion of these. Haladyna and Downing (2004) state that research is starting to recognize that not only the construct itself but also construct irrelevant variance should be included into this research. This means that social consequences, how the test is experienced and the uses of the test should be looked into. Other research has also advocated for inclusion of test takers’ perspectives (i.e. Moss, Girard & Haniford, 2006; Fox & Cheng, 2007). However, up until now, Cheng & DeLuca (2011) is one of the few research that actually included test takers’ perspectives on the testing experience and the relation between this experience and (perceptions of) test validation and use. They did a qualitative study in which they looked into 59 reports written by test takers after a range of tests in English. The themes incorporated in the study were the test’s administration and testing conditions (bias), timing (time pressure), test structure and content, scoring effects (i.e. lack of feedback), preparation & test-taking strategies, test purpose, psychological factors (i.e. stress, anxiety), external factors, and test consequences (i.e. stakes of the test). Their study showed that test takers felt that the themes impacted their test results, thus impacting the validity of the tests. This impact was complex and multifaceted. Therefore, stakeholders’ opinions can provide valuable insights into validity and should thus be included into this research. Moreover, Ryan (2002) believes that especially in high-stakes assessments certain stakeholders like test takers or teachers view the test quite distinctly from the government or the like. Including these perspectives in the validation process may therefore provide useful information in the identification of strong or weak points in the test’s interpretation.

The studies mentioned above report mixed findings. It is clear from these, however, that washback exists and should be taken into account when looking into validity. Moreover, not all studies have used the same methodology to look into washback. In this study, washback will be looked into through qualitative interviews with both teachers and test takers. One class observation is done as well.
Validity in computerised tests

The first automatically scored test was reported by Page (1966). She looked into an automatically scored test for essays, which was set up to lessen teachers’ work-load. Due to a lack of computers these tests were not an option yet. As more computers became available in the 1990s, more automatically scored tests were developed. Nowadays, most automated scoring technologies are used in writing tasks (Fulcher, 2010).

However, the new medium does not mean that computerised tests are essentially new or improved. That is, computerised tests are frequently used to facilitate current test administration, whilst there are a number of possible innovative options available for computerised tests (Chalhoub-Deville, 2002). These options can be found in adaptive language testing, for example. These tests were named Type I (traditional tests administered via computer) and Type II (innovative tests which use the computer to its full potential) by Maddux (1986). Chalhoub-Deville (2002) looked into the range of tests available at that time and found that most tests can be defined as Type I tests. This means that although computerised tests may be thought of as innovative and better than traditional tests, they may very likely not be.

Moreover, the development and use of computerised tests is received with scepticism. On the one hand, the introduction of such tests would relieve the burden on teachers, produce impartial scores and reduce costs in large-scale assessments. On the other hand, there is suspicion about the validity of these tests. Therefore, the validity should be looked into (Clauser et al., 2002).

The validity of automated scoring technologies has mainly focused on checking the tests’ reliability by means of comparing computer scores with human scores, looking into patterns of these two scores in combination with other measurements (i.e. Bernstein et al., 2010), and understanding the scoring technologies (Yang et al. 2002). However, a mere check of the reliability of automated tests is not enough to validate them. Other aspects of validity should be looked into as well (Yang et al. 2002). Fulcher (2010) adds that even though high correlations between human-raters and system scores (the ‘gold standard’) were found, it does not necessarily mean that both are measuring the same construct or ability. Humans can focus on effective communication whilst a system can focus on pronunciation. Moreover, these automated systems do not provide feedback like humans do. Therefore, correlations are insufficient to claim validity (Bejar & Bennett, 1997). Chapelle (2001) states that in order to be able to infer from a (computer) test there has to be some degree of authenticity between what is tested and what the target language use is, as there inevitably is a difference between what is tested and what is actually used.
These research show that automatically evaluated scoring systems are usually only evaluated on reliability aspects. However, there are clearly other validity problems in these. Some validity issues have been raised in the past on the Phonepass test, a test which is the precursor of the TGN. These issues are listed next.

**Validity of the Phonepass.** The TGN is an adaptation of the Phonepass test (created by Ordinate, US), which includes less subtests, but is based on the same principles. Several researchers have looked into this test’s validity, although most focused on its reliability (de Jong & Bernstein, 2001; Bernstein, de Jong, Pison & Townshend, 2000). The reliability of the Phonepass in comparison with human-raters proves to be good.

Becker & Ribeiro (2005) looked into the test’s validity by assessing documents produced by reviewing documents which Ordinate had written on the Phonepass’ validity. They found that the test was indeed highly reliable but lacked predictive values for longitudinal performance and is not in line with what linguists call ‘fluency’.

Chun (2009) looked into the test’s authenticity and concluded that the test fails in being authentic. In fact, he claims that the abilities shown in the test do not predict nor equal speaking abilities shown in everyday-life. The telephone and actual listening and speaking involve different cognitive functions. His advice therefore is that the test needs major revisions. His review met criticism of Downey et al. (2008), considering his inaccurate application of the Bachman & Palmer (1996) model. Downey et al. (2008) state that a test cannot be considered invalid based on a low degree of authenticity (which he denies as well). Consequently, Downey et al. (2008) applied the model to the PhonePass test, although they focused mostly on reliability (even though they claim this is construct validity) and leave out the impact and interactiveness. In short, neither party applied the model correctly.

In her article, Xi (2010) discusses the difficulty automated scoring technologies have with the accurate recognition of accented speech. Moreover, different intonation and stress patterns are rather difficult to recognize for these programs. This is especially the case in spontaneous speech. She comments on the Phonepass tests and the use of highly restricted tasks where no spontaneous speech is allowed. Xi (2010) states that this approach still under-represents the construct of speaking proficiency. It is very limited in its measurement. Moreover,

*Given the limitations of current speech scoring technologies, if they were used to score an assessment to support high-stakes decisions, the assessment may be vulnerable to new types of test-taking or cheating strategies. Test takers would be prompted to (...) speak to the less sophisticated machine grader to try to achieve*
undeserved high scores. This would negatively impact the trustworthiness of the scores. Further, if the test tasks or the automated scoring model under- or misrepresent the construct of interest, test takers may be led to place an inappropriate focus on wrong skills or to omit important skills in their test preparation. An automated test that under- or misrepresents the construct may also bring about negative washback effects on teaching and learning and compromise the credibility of the test program.' (Xi, 2010: 294).

The previous mentioned research shows that the tests are highly reliable but lack in authenticity and predictive longitudinal performance, although the Phonepass does not exactly measure the same constructs as the Dutch TGN does. In short, the effects an automated speech scoring system can have on the validity of the test are substantial.

It is clear from the above overview that the way validity was looked into in automated scoring systems mainly focused on how reliable the scores are. Moreover, the precursor of the TGN has received abundant criticism as well. Therefore, the TGN will be looked into with the previously explained model created by Bachman & Palmer (1996). The model involves all facets of validity and has only been applied to automated scoring technologies by Chapelle (2001). His research involved two reading tests (of which one adaptive reading test), a listening and a writing test. However, he merely showed how the framework could be applied to automated tests. He did not make any judgements about the tests.

Thus, although the model created by Bachman & Palmer (1996) exists and can be applied to automatically scored tests, this has not been done much yet. In this thesis the model will be applied to the TGN as it is clear from the above that there may be validity issues with the TGN.

**Statement of purpose**

It is clear from the above literature review that the validity of this high-stakes test is likely to be in danger. The consequences of the test can be substantial as well. It is vital that this test is valid since the TGN is currently (since January 2013) the only measurement of spoken language ability in the Civil Integration Exam. Therefore, the main purpose of this study is to gain insight into the validity or usefulness of the TGN as used in the Civic Integration Exam. Due to its obvious subjectivity, the evaluation of usefulness of the TGN is reviewed by a panel of experts. Moreover, this thesis combines the model with the influence of the test on teaching and learning in the preparation of the test (washback). The study combines insights from theory with valuable information gathered in the field. Next, the methods in researching the validity used are reported.
Methodology

This section will describe how the validity of the TGN will be checked by an analysis based on the model of Bachman & Palmer (1996), for which evidence had to be collected. First the subjects who participated in the interviews and the panel of experts will be described. The materials that were used in the analysis of the TGN, the interviews and the panel-review are discussed next. These are followed by the procedures in all three stages of this research. Every sub-section describes this order: analysis according to the Bachman & Palmer (1996) model, the interviews, and the panel-review.

Subjects

This section describes the subjects used in the analysis, interviews, and the panel-review.

Analysis Bachman & Palmer (1996) model. No subjects were used in this stage of the research.

Interviews. Interviews were conducted with stakeholders in the field in order to get insight into the social aspects of using the TGN. In this way, both the test's impact and its potential washback could be looked into. First, the test takers will be described, followed by the teachers.

Test takers. In total, four test takers who passed the naturalisation exam before 2013 were interviewed. The test takers were found through a post on Facebook which was forwarded via the medium. Two out of four test takers are Chinese, one is from Singapore and one is from Mexico. All participants came here to join their partners and are now studying at universities in the Netherlands. In total, three females and one male were interviewed. Of these people, one has proceeded by doing the higher leveled (B1) ‘Staatsexamen I’ (State exam I), the other three are planning to finish that this summer. Note that the only person who was interviewed in Dutch is the female who completed the B1 exam. The others preferred English.

Teachers. In total, six teachers from five different institutions were interviewed. All teachers are female and have been teaching for several years. Two teachers work together in a small company (they are the only teachers there) with classes up to 15 students. One teacher teaches private lessons, another teaches groups of up to 30 students in a large institution. The final two teachers work at institutions for vocational education with groups of 20-25 students.
Panel-review. Once the analysis was completed, a panel of three experts in the field of Dutch as a second language and testing was set up. Two of the experts are involved in the Board of Examinations (Commissie voor Examens, CvE) in the Netherlands which means that they have substantial knowledge about what exams should look like. All members are experts in the field of testing and Dutch as an L2.

Materials
This section describes the materials used in the analysis, interviews, and the panel-review.

Analysis Bachman & Palmer (1996) model. The official TGN was not made available for this research. Therefore, other official documents and the practice version of the test had to be used to collect evidence of usefulness. First, the construct and the test components of the TGN were reviewed using the justification of the test made by its developers: the Verantwoording Toets Gesproken Nederlands (Kerkhoff, Poelmans, de Jong & Lenning, 2005). This document was complemented with an article written by the test-developers on the TGN used at embassies to check the Dutch on the A1 level of immigrants-to-be (de Jong, Lenning, Kerkhoff & Poelmans, 2009). Since the first document should explain the exact construct, the decisions that were made in the development of the test and other important facts about the test, this document was considered the most important and used accordingly.

Interviews. Underneath the materials used in the interviews with test takers and teachers are discussed.

Test takers. The interviews were used to get insight into the impact of the test on the test takers (and their surroundings) and the washback of testing on their learning of Dutch. Questions were made accordingly. Again, the true intention of the research was not immediately made available to the test takers. Therefore, questions on another test on the Dutch society were asked as well as questions on their overall sentiments of the exam. Every interview started with a few basic simple questions on their home-country and their reason for coming to the Netherlands. Only open-ended questions were asked next. These can be found in the appendix. It should be noted that the questions were used as a guideline. Whenever interesting information was given more detailed questions were asked.

Teachers. The interviews were used to get insight into the consequential aspects of the use of the TGN. Besides the impact of the test on teachers, the washback on teaching was looked into. The teachers did not know exactly what the goal of the interview was in
advance. Teachers were gathered by asking them to be interviewed on how they prepare for the entire exam. Again, the reasoning for this was to decrease the teacher’s bias and to gather as honest answers as possible. Questions were made accordingly. The interviews with the teachers started off with questions on how long they have been teaching for and how they entered the profession. The questions (which were used as guidelines) can be found in the appendix. Again, whenever interesting points were discussed the researcher asked more detailed questions on these.

Panel-review. Once the analysis according to the Bachman & Palmer (1996) model was completed, the panel was given the analysis together with a form on which all qualities and their sub-questions were listed. This form included boxes which would be ticked ‘agree’ or ‘disagree’. Another box offered the possibility to explain why the member did or did not agree. The form can be found in the appendix.

Procedures
This section describes the procedures followed in the analysis, interviews, and the panel-review.

Analysis Bachman & Palmer (1996) model. Before the actual construct was analysed an overview was made on what the test claims it tests. Obviously the construct needs to be compared with the intentions in order to get insight into the usefulness of the construct. Then, the construct was reviewed by looking into the articles and books referred to in the justification. Moreover, for every claim made in the construct it was checked whether there was a reference present and whether the reference was paraphrased and used correctly. In this way, the theoretical basis underlying the TGN could be analysed. When a claim was made in the justification that could not be checked within the two documents, the practice test was used. This test is supposed to be a smaller, yet comparable version of the TGN. Finally, the criticism on the TGN’s construct was listed and used for the model of Bachman & Palmer. How the model was construed with the evidence will be explained later.

Next, a comparison was made between the A2 level of the CEFR and the justification of the test. Therefore, the requirements for both the A1 and A2 level on listening and speaking were gathered and compared with the different subtests of the TGN: repetition, short answers & antonyms. The final subtest (retelling of stories) was left out of the analysis since this part is not scored. Since the official test was not made available for the current purposes, the practice-test and examples in the justification were used in the comparison.
Interviews. This section describes the procedures in interviews with both test takers and teachers.

Test takers. The interviews were conducted via Skype and recorded with a dictaphone. After administration the interviews were summarised and completed with information (about what happened in class, for example) given by test takers after the actual interview. The extraction of information was maximised in three ways: First, participants were told the interview would be about the entire naturalisation exam without mentioning the purposes (looking into the validity of the TGN). This was done to limit the bias. Moreover, they knew that all information given would be made anonymous. Finally, all questions (besides questions into basic information) were open-ended as to elicit as much information as possible.

Teachers. Teachers were gathered via the Dutch website www.blikopwerk.nl which lists all institutions offering courses to get to the A2 level of Dutch. This website checks the quality of language institutions and rates them accordingly. Due to time restrictions only institutions which were within two hours travelling from Utrecht were e-mailed. In the e-mail, teachers were asked to inform the researcher about the way they prepared their students for the A2 levelled exam. Also, they were told that all gathered data would be reported on anonymously.

When teachers were interested an appointment was made to interview them at a location of their preference. All interviews were conducted in their work offices in one-on-one situations. A dictaphone was used to record the interviews. After administration interviews were summarised and completed with information given after the dictaphone had stopped recording. That is, after the actual interview teachers were told what the actual goal of the research was (the validity of the TGN). This always resulted in useful material which was written down on paper by the researcher. Due to the degree of importance and relevance of this material, it was decided that this information was added to the summary.

Observation. One language institution offered the opportunity to attend one of their classes. Although it originally was not within the scope of this thesis to observe classes and add these to the assessment of usefulness, the opportunity was accepted. Previous research indicated that these observations could add to the information on washback since teachers might state they do not teach to the test but actually do. During the observation notes were made on the impact and washback of the test. These were then added to the interview held at this institution.

Panel-review. The three members were asked to join the panel via e-mail. Once they agreed, the analysis of the TGN and the previously discussed form were sent via e-mail. The panel was instructed to limit their time spent on evaluating the analysis and to tick the boxes
‘agree’ or ‘disagree’ in the boxes. All members sent their evaluation back via e-mail. Once these were obtained, the results were analysed and incorporated in this thesis.

The next section consequently lists the results of the analysis, interviews, and the panel-review.

**Results**

This section describes the results of the analysis of the TGN, the interviews and the final evaluation of usefulness consequently. Finally, this chapter ends with the results from the panel-review.

**Analysis TGN**

An analysis of the TGN was carried out before turning to the TGN’s actual validity in the next section. First, the request for a test by the Dutch government is explained, followed by the test’s construct and administration. Then the claims the authors make in the test’s justification are reviewed followed by a comparison of the claims with the CEF.

**Development of the TGN.** In 2003 a request was made by the Ministry of Justice to set up an oral exam linked to the CEF for both immigrants-to-be and immigrants already living in the Netherlands. This request was picked up by main contractor CINOP (NL), who hired Language Testing Services (NL) and Ordinate Corporation (USA) (de Jong et al., 2009). The latter company provided CINOP with the testing system.

The TGN was originally set up to test whether the listening and speaking skills of immigrants-to-be are at the A1- CEF level (below beginner level: also referred to as tourist) and could also be used as examination for naturalisation of immigrants already in the Netherlands at the A2 level (waystage or intermediate level) (de Jong et al., 2009). Both exams are the same, but the caesura is different. The next section will describe the test’s administration and construct.

**Administration & construct TGN.** The test is administered via telephone and consists of four subtests: sentence repetition, giving short answers to questions, naming antonyms, and retelling stories. The final subtest is said to be used to validate the test and is not used to decide if the subject’s Dutch is at the A2 level. The answers given to the first three subtests are then analysed and scored by the automated speech analysis system of Ordinate Corporation which is trained to do this through many utterances. In total the exam lasts 12 minutes. Following table (1) (which shows the number of items per subtest with an example), all subtests will be explained.
As can be seen from the table, before each subtest a non-scored practice-item is given. The subtest sentence repetition is divided into two parts (before and after short answers) and only one practice-item is given at the beginning of the test. This subtest starts with short sentences of two words minimum up to longer ones of thirteen words maximum. The part in which stories are to be repeated does not contain any practice-items since this part is not scored. A practice-test containing 30 items of the TGN is available via the Dutch telephone-number +3188-7890123, which gives an indication of items in the actual TGN. This version can be found in the appendix.

Each task is read at a natural pace by L1 speakers of Dutch: men and women speaking different regional accents. In order to only test oral proficiency the construct controls for knowledge of the world, level of education and cultural identity (Kerkhoff et al., 2005). Put differently, the construct is based on the idea that a child aged 12 without knowledge of the world should be able to answer the questions and repeat the sentences (de Jong et al., 2009).

The test-items were constructed and then compared with the Corpus Gesproken Nederlands (CGN; Corpus of Spoken Dutch) to control for the level of vocabulary. Finally, experts in the field of Dutch as a second language reviewed the produced items. In the beginning, the pre-test consisted of 2131 items. Since then, more items have been created and added to the item-bank managed by CINOP. The exact number is unclear.

The scoring of answers is done automatically by the Ordinate system, which was trained by speech of both L1 and L2 speakers of Dutch. The system judges answers in two ways: on what is said (based on sentences and vocabulary) and how this is said (fluency and pronunciation). Basically, the system judges the content and the quality of the content,

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Number of items</th>
<th>Number of scored items</th>
<th>Aspect</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence repetition</td>
<td>2 x 12 = 24</td>
<td>23</td>
<td>Pronunciation, fluency &amp; syntax</td>
<td>Daar heb ik nog nooit van gehoord. (I have never heard about that)</td>
</tr>
<tr>
<td>Short answers</td>
<td>14</td>
<td>13</td>
<td>Vocabulary</td>
<td>Kun je rijst eten of drinken? (Can you eat or drink rice?)</td>
</tr>
<tr>
<td>Antonyms</td>
<td>10</td>
<td>9</td>
<td>Vocabulary</td>
<td>Ochtend (morning)</td>
</tr>
<tr>
<td>Repeating stories</td>
<td>2</td>
<td>-</td>
<td>(Validity)</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>50</strong></td>
<td><strong>45</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 - Subtests TGN, number or (scored) items, test aspects & examples
respectively. Both aspects make up fifty percent of the score, as both can cause miscommunication.

According to the test-developers, the construct is based on psycholinguistic models of language (Kerkhoff, *et al.*, 2005). The goal of the test is to measure ‘(…) the facility with which candidates are able to track what is said, extract meaning in real time, and formulate and produce relevant, intelligible responses, at a conversational pace.’(de Jong *et al.*, 2009: 43). This goal is based on Levelt’s (1989) model of conversation.

**Literature review of justification TGN.** To justify the TGN, the test-developers claim that sentence repetition can be used to get an insight into the participants’ working memory (WM) as people can remember sentences of fifteen to twenty words if these words are meaningfully connected. They fail to do this after seven words if this connection is not present for them. This claim is based on research by Baddely (1986; 2000) and Poelmans (2003). However, when looking at the research by the latter, this claim is not researched at all but is taken from Baddely (1986; 2000).

Moreover, when looking into the TGN’s official practice test, the average number of words in the sentences of the repetition task is 5.86 words (N=14, Min-Max=3-9, SD=1.75). Still, when leaving out the two outliers (three & nine), the average sentence-length is 5.92. A mere five out of fourteen (36%) sentences contain seven or more words: sentence-lengths which, as stated by Baddely (1986; 2000), can be remembered without the need for a meaningful connection between the words. This means that on average participants do not need to understand the sentence in order to repeat or imitate them correctly. If the practice-version of the test really is a reflection of the actual test (which the test-developers claim it is), then more than half of the items in the repetition task cannot give insight into language proficiency. These merely show how well a person can imitate. Scoring these would not make sense. However, this is done in the TGN.

Furthermore, on page 24 the test-developers justify the use of the repetition tasks by saying that sentence imitation is not uncommon in conversing and interacting with other people, for which they refer to van Baaren *et al.* (2003). However, the article does not make this claim at all. In fact, the article describes an experiment in which a waitress either repeated the order or merely confirmed it. This is not the same as considering repetition a normal aspect of communication. Furthermore, none of the research mentioned in the article by van Baaren *et al.* (2003) claims that mimicking of facial expressions or syntax is quite common; repeating complete sentences is not. Kerkhoff *et al.* (2005) also claim that the level of correctness and fluency gives information about the automation of their linguistic processes in Dutch – this claim, however, is not based on any references.
The researchers (idem) also state that the normal interaction time between the coding and realisation of an answer takes up 40 ms (Van Turennout, Hagoort & Brown, 1998). Turn-taking in normal interaction takes up about 500 ms (Bull & Aylet, 1998). These two are combined in a claim that in order to participate in a conversation, certain processes (retrieving information from mental lexicon, construing sentences, and producing these without having to think about it) need to be automated (Kerkhoff et al. 2005, based on Cutler, 2003; Jescheniak, Hahne & Schrievers, 2003; Levelt, 2001). The TGN then measures to what extent the processes (and combinations of those processes) mentioned by Levelt (2001) are carried out automatically (Kerkhoff et al., 2005). By looking at these processes the test-developers think that they can predict the ease with which the participant can join in on actual conversations. It should be noted that their way of testing these processes in the TGN is not supported by any reference whatsoever.

The subtests questions and antonyms are not based on any literature. Although intuitively it seems appropriate that these tests are included, as one would expect references to previously conducted research into these aspects. Considering this is not the case, it remains questionable whether these aspects are useful in this test.

Although the test-developers claim that they have looked into the TGN’s validity, the test’s justification mostly shows measures of reliability. They describe validity as the need for the system to be accurate. This is tested with a comparison with human-raters which were trained with the CEF levels. An overall correlation of .93 was found between the automatic scoring system and the human-raters. However, this is still an indication of how reliable the system is. It does not give information on the construct validity, authenticity, interactiveness, impact or practicality of the TGN. Therefore, the next chapter will evaluate the usefulness of the TGN in the light of these aspects of validity. First, the next section will compare the TGN with the CEF.

**Comparison of TGN with CEF.** Before starting up the construction of the test, the Ministry of Justice explicitly mentioned that the test tasks should be related to the CEF (De Jong, 2009). The test was constructed so it could measure the range of no facility in Dutch to perfect facility (Kerkhoff et al., 2009). Then, the A2 level was chosen for proficiency in the Civic Integration Exam in the Netherlands. One could reasonably expect that every subtest in the TGN relates to at least one aspect in the CEF A2 level (Council of Europe, 2001). The A2 level will be linked to the TGN.

The Council of Europe compiled a document which states that a citizen of Europe should preferably be a multilingual to overcome possible communication barriers. Moreover, the document was set up to ‘provide a common basis for the elaboration of language
syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe’ (2001:1). It describes six levels (A1, A2, B1, B2, C1 and C2; the latter being the highest) of language proficiency. These levels are explained in more detail for the four skills - reading, writing, speaking & listening - and spoken interaction of which only listening and spoken production will be compared with the subtests as the TGN only wishes to get insight into these. Table (2) shows these skills for the A1 and A2 level.

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Listening</strong></td>
<td>I can recognise familiar and very basic phrases concerning myself, my family and immediate concrete surroundings when people speak slowly and clearly.</td>
<td>I can understand phrases and the highest frequency vocabulary related to areas of most immediate personal relevance (e.g. very basic personal and family information, shopping, local area, employment). I can catch the main point in short, clear, simple messages and announcements.</td>
</tr>
<tr>
<td><strong>Spoken Production</strong></td>
<td>I can use simple phrases and sentences to describe where I live and people I know.</td>
<td>I can use a series of phrases and sentences to describe in simple terms my family and other people, living conditions, my educational background and my present or most recent job.</td>
</tr>
</tbody>
</table>

Table 2 - CEF description of listening and spoken production levels A1 and A2 (Council of Europe, 2001: 26)

The first subtest, repetition, is claimed to require the understanding and reproduction of the sentences. However, the justification does not claim any relation of these skills to the CEF. This is most likely due to it not being present. The longer sentences are most likely too difficult to understand for subjects at the A2 level; they only need to understand shorter sentences. When we turn to the spoken production in this subtest (which is in fact repetition) we can see that a person on the A2 level needs to be able to describe subjects which are close to him/her or describe him/her. However, when we look at the practice version of the TGN hardly any sentences can be found which include any of the subjects mentioned in table (2). In short, the first subtest does not seem to conform to any of the skills described either on the A1 or A2 level.

The second subtest, short answers to questions, requires the understanding of the sentence and the possible two answers to the question. This subtest is targeted on measuring the vocabulary of the subjects. Considering the use of the CGN (from which the test-developers used the most frequent words) in the development of the test it can be assumed that subjects should be able to know these words. In fact, the CEF requires subjects to know the highest frequency vocabulary. However, the ability to answer questions is not required in the CEF.
Giving antonyms is the final scored subtest in the TGN. Here a subject needs to understand the word given and state the opposite word. Again, this ability has not been defined in the CEF.

In conclusion, the sub-tests are difficult to link to the CEF. Besides a literature-review, interviews were conducted. These are reported below.

**Interviews**

Other evidence into the validity of the TGN was gathered by interviewing stakeholders in the field. In this thesis four test takers and five teachers from different language institutions were interviewed. Summaries of these interviews can be found below. Interviews with the first stakeholders will be discussed followed by interviews with the latter group.

**Test takers.** The test takers are from Singapore, China and Mexico. Summaries of the interviews are reported.

**Male from Singapore.** He came to the Netherlands from Singapore because of his partner and has been here a little over a year now. Prior to coming to the Netherlands he already had to pass the TGN at the embassy. He was helped by his partner who flew down to give him some books to practice. He was then allowed to come to the Netherlands where he had the opportunity to pass the Civic Integration Exam within three years. He passed the entire exam in the beginning of 2013.

He was exempted from the TGN as he scored higher than 37 points when taking the test at the Embassy. However, he did not always understand what was asked: ‘O my god, that was just crazy. It was very fast. And to be honest, I did not know what I was doing’. His strategy for the repetition tasks was to remember the first and final few words as memorising the whole sentence was impossible. He just rattled when he did not know the words. This strategy obviously worked for him. The book he got from his partner was helpful as well in learning certain tricks. He passed the test and got A2 immediately, even prior to coming to the Netherlands.

He understands the ideas behind the test but points out that it only focussed on pronunciation and that the speed of the test was too fast. He believes that the speed of the test was too fast, even for more able speakers of Dutch.

Preparing for the test was stressful; a lot depends on passing the test. It was also very expensive. The impact on him and his partner was rather big as well. His partner helped in the preparation but both got frustrated with the process. However, he feels it is important and useful that the Netherlands have integrated such an exam into the process of becoming
Dutch. Although almost everyone can speak English, he feels more in place now that he speaks some Dutch. However, he does not feel that his Dutch is very good and thus does not dare to speak a lot. He will continue with STEXI soon.

**Female from China (I).** When doing her PhD in the Netherlands she fell in love with a Dutch man, married him and wanted to stay in the Netherlands permanently. Therefore, she had to pass the Civic Integration Exam which she did. She did not take an actual course, just a volunteer class. She learned most because she has been here for five years and she knows someone who had a book and wanted to help her. She watched the news as well which was helpful. She found that the speed of the TGN was rather fast, even for Dutch people. She prepared by speaking with Dutch people, she did not really use any books or materials as she did not know these were available. For her the exam was not that difficult as she had already spent five years here. The exam was not very useful for her as she knew she could already speak Dutch. It might be useful for people who do not know a lot of Dutch but she knows of people who have forgotten everything they had learned already.

**Female from China (II).** This Chinese girl got the opportunity to do the ‘inburgeringsexamen’. She thought it was a nice chance to learn Dutch. At the same time she was also learning English. This made it a lot easier for her. On top of that she believes she is a fast learner. She took a course and learned from books. These helped her to learn Dutch but she did not pass the TGN in her first attempt. She knows a lot of people who have had the same problem. Overall she considered the TGN difficult, especially the repeating part. For the other things she could just study. The repeating part went rather fast and she could not understand the sentences immediately. She usually said the first few words to score enough points. She would have preferred it if she could have asked the computer to slow down a bit or repeat it for her, just like she would ask people in actual conversations. Moreover, she considers the TGN to be old-fashioned. She hardly ever speaks Dutch now. She feels that she cannot do it very well.

**Female from Mexico.** This Mexican woman has been living here for three years and is studying at a university. After living in Flanders, Belgium for a year where she worked as an au-pair and took classes of Dutch, she fell in love with a Dutch man and decided to live in the Netherlands. She took the TGN at the embassy where she passed the test on the A2 level. Therefore, she was exempted from the TGN in the Netherlands. She practised for the test with her boyfriend and focussed on speaking without much context or body-language: she tried to speak Dutch on the phone. The exam had a rather big impact on their relationship. Both knew how much depended on the exam. It caused tensions but made her
speak better Dutch. She did take a course after she completed the A2 levelled exam since she thought her level of Dutch was not good enough. Finally, she completed STEXI.

She thinks the test is rather stupid and simplistic. She does not believe the test is useful for her. She knows about the culture in the Netherlands. It might be useful for cultures which are very different from the Dutch society. The antonyms were rather difficult for her as she only practised using free on-line material. Besides that, she practised with her boyfriend. However, her living in Flanders for a year had helped. She still learned to the test but claims it was just easy for her.

**Summary interviews test takers.** All test takers have recently passed the Civic Integration Exam and were eager to learn Dutch. All of them are now doing or have just finished university, either in the Netherlands or in their home countries. Therefore, it can be concluded that all test takers are intelligent. They indicate that they most likely passed the test because their ability to study.

The speed of the test, especially in the repetition task, was considered very fast which caused stress and not much understanding. They would all like to see the speed slowed down a bit. Two out of three test takers used a trick (only repeating the first few words). All of them learned the antonyms from heart and two out of three studied for the questions. The remaining participant claims she was already rather fluent and had no trouble answering the questions. The same test takers report to find it difficult to speak Dutch. They learned mostly for the test and are eager now to do the State Exam I in so they are more able and willing to speak Dutch.

**Teachers.** Summaries of interviews conducted with five teachers from four different language institutions are reported next. The sizes of the institutions differs per teacher.

**Private classes, one teacher.** This female teacher has been giving Dutch classes for six years. As classes are diminishing she mostly teaches private classes now. This is also due to time restraints. People need to learn Dutch as fast as possible. You can do more in one-on-one classes. She prepares her students with practice-tests and lists of antonyms. When people know these they can pass more easily. She considers the TGN as a trick you can either do or cannot do. Some people will make it and some won’t. Some people are on the A2 level and continuously fail the exam, whilst some bluff their way through it. This is a small number of students but still relevant because it is possible to do so. Moreover, the teacher thinks that the entire skill of spoken language ability is not tested, just vocabulary and pronunciation and that this is not fair. Moreover, she does not believe that a computer
can actually hear the difference between what is right and what is wrong. The aspect of non-verbal communication is left out as well, although it is very helpful in real communication.

Further to this, in the adjusted Civic Integration Exam of 2012 exam, the TGN is the only test that looks into spoken language ability. Previously there was also the portfolio for which people had to speak with other people and prove what they were able to do. Although it was difficult for some, people learned to speak Dutch this way.

This teacher agrees with the recent discussion on the test and hopes the Ministry will work something out so more aspects of spoken language will be involved in the exam. She believes that the TGN can exist alongside an exam that tests spoken language ability as a whole, preferably through actual communication with a person.

**Very small company, classes of 10-12 people.** This small company was set up by two teachers in 2007. They started in groups of 20-25 people but lowered the group-size to 10-12 people. Every person gets his or her own trajectory and personal attention. In this way no student gets bored or feels it is too difficult. The goal of the courses is not necessarily to pass the exam but to get the people to the level of A2 (preferably higher) so they can find a job and cope in the Netherlands. They acknowledge that they know many teachers from other companies or institutions who only teach to the tests. However, they believe that their method has gotten them to their 100% success rate.

Although the teachers claim they do not teach to the test they focus on pronunciation a lot and let their students practice opposite words and teach them to answer questions in short. They admit that if they do not focus on it at all the students will not pass the exam. Ultimately, this is the goal of course. This only happens in the final few weeks. Both teachers state that the TGN is not a good test and would like to see a more authentic and levelled test, especially now that the practical exam has been eliminated. They understand why the TGN is used but not why this is the only way of testing speaking abilities.

**Large company, classes of 25 people.** This woman has a lot of experience in teaching Dutch as a second language, working as both a teacher and head of the department since 1996. They offer naturalisation courses but this is a very small part of the company. After an intake people are put into classes of 20-25 people. First, every student gets approximately 30 hours of classes four times before taking the exam. The level of Dutch needs to be improved first before preparing for the actual exam. They have books and other materials for the TGN to prepare people for the test. They also train them on the different accents as these are perceived as difficult by people who take the test.
Although she finds the TGN absolutely terrible, she understands why this test was chosen. However, she cannot cope with the statements that the TGN tests spoken language ability since it does not. Most of it is repetition of sentences you do not necessarily have to understand. You can just repeat a few words and then you pass. When people are having a lot of trouble with the test, they are taught this trick. The rest they can be learned from the top of their heads. She does note that the questions are ridiculous: you do not take these people seriously by asking these questions. Overall, she believes there must be a better way of testing these people and treating them with more respect.

**State funded institution for vocational education (I).** This teacher has been teaching Dutch for over 20 years. She explains that they get people to the A1 level first and then start focusing on the exam. Without asking she states that preparing for the TGN is an enormous problem. People prepare themselves for it with books and they used to have a speech pathologist assisting them. She said that during the classes they train them for the TGN but see that remembering and repeating sentences is strenuous for former illiterate people. When you come from a country where recognising sentences is normal or when you posses study skills needed for repeating sentences then you can pass the exam. People who are or were illiterate cannot pass due to their lower working memory skills. The test is not motivating for people. There is a lot at stake.

Moreover, she feels that the test is not good due to the few things it measures. It claims to measure spoken language abilities but in fact it only touches the surface of it. It is not fair that the test-developers claim that it tests spoken language ability. She believes the name ‘parrot-exam’ fits the test perfectly. For most things you can just study but the speed of the repetition is so fast that people can barely breath. She believes that the repetition part should change.

If she could change the exam she would most likely reinstate the portfolio. When it was introduced she did not see the use of it but after a while (and especially now that it is gone) she sees the use of it: It made people talk. These people also found it useful, they learned a lot with it. She does not believe the naturalisation exam should be about using tricks but about participating in the country with the people. That was the original idea of the ‘inburgeringsexamen’: The TGN surpasses this idea.

**State funded institution for vocational education (II).** This teacher has been teaching since 1999. After an intake people start the course in a folder of their level. They practice a lot for the TGN from the beginning. They have practice tests and prepare the test in groups as well. Every once in a while she has to teach people a trick which entails
repeating the first few words in longer sentences. She insists that people practice this every day as it can be rather strenuous if they start to prepare only weeks before the test.

She feels that with the omission of the portfolio, the exam has lost a vital part. Now there is only the TGN testing spoken language abilities (or claims to test this ability). ‘I do not believe that the TGN is a good tool for testing spoken language abilities on its own. With the portfolio people actually had to speak which, in turn, they dared to do. The TGN does not make people speak. No matter how difficult the portfolio was, people actually learned something.’ With only the TGN left, she fears that she has to teach grammar-classes again in order for people to pass the test. In short, she agrees with the recent discussion in the field about the TGN. If there was anything she could change in the ‘inburgeringsexamen’ it would be adding a part which actually tests spoken language abilities through, for example, an actual conversation.

**Summary interviews language institutions.** All five language institutions have been working with Dutch as a second language for a while. Although the way of teaching and the group-sizes are rather distinct, all teachers have more or less the same ideas about the TGN. None of them agree with the TGN being the only instrument that tests (aspects) of spoken language abilities. Moreover, not a single teacher spoken with was positive about the test.

Although all teachers who were interviewed agree that teaching to the test is not a good solution, all of them do this at a particular stage in the process. They feel that otherwise the test cannot be passed even though the time spent on teaching to the test is better spent differently. What is more, every teacher has had to teach several students a trick to pass the exam.

When asked the question ‘How would you improve the exam?’, all teachers wished to see a more communicative approach to testing. Four out of five teachers would therefore like to see the portfolio reinstated. This offered people the chance to actually speak with other Dutch speaking people.

**Evaluation of usefulness**

The six types of validity are explained and applied to the TGN below. All questions below were taken from Bachman & Palmer (1996: 150-155).

**Reliability.** Reliability scores were checked by Kerkhoff, *et al.* (2005) for 139 participants by means of split-half reliability. Overall, the reliability scores on the TGN were .94 for the complete pre-test. Pronunciation and fluency both scored .89, lexicon scored .73 and syntax
.93. These reliability scores thus differed. Later, reliability showed to be .91. The reliability was also checked with scores of humans who were trained to rate the answers given by participants the same way as the system does. These correlations ranged from .80 to .94.

1. To what extent do characteristics of the test setting vary from one administration of the test to another?

Quality satisfied - There is no variation in setting. The test is administered via telephone. The noise-levels may differ on locations which can affect the reliability.

2. To what extent do characteristics of the test rubric vary in an unmotivated way from one part of the test to another, or on different forms of the test?

Quality completely satisfied - Instructions or rubrics are the same for everyone which can be attributed to the use of the automated system.

3. To what extent do characteristics of the test input vary in an unmotivated way from one part of the test to another, from one task to another, and on different forms of the test?

Quality more or less satisfied - The test’s input differs per person as items are randomly chosen from a database of thousands of items.

4. To what extent do characteristics of the expected response vary in an unmotivated way from one part of the test to another, or on different forms of the test?

Quality completely satisfied - The expected (i.e. correct) answers do not change and have not changed in the past years.

5. To what extent do characteristics of the relationship between input and response vary in an unmotivated way from one part of the test to another, or on different forms of the test?

Quality completely satisfied - The test is the same for everyone at any time. The difference in parts of the test have been motivated. This (and other parts of the test’s reliability) has been evaluated deeply (see Kerkhoff et al., 2005).

Construct validity.

1. Is the language ability construct for this test clearly and unambiguously defined?

Quality more or less satisfied - The language ability has been explained clearly: it seeks to see how fast a candidate can understand the meaning of sentences and can reply to them
VALIDITY OF THE TGN

(Kerkhoff, et al., 2005). However, why the different subtests were chosen is not clear at all. This should have been added and explained in depth.

2. *Is the language ability construct for the test relevant to the purpose of the test?*  
Quality not satisfied - The test’s purpose is to get insight into the listening and speaking ability (which was asked by the Dutch Ministry of Justice) as measured by *‘the facility with which candidates are able to track what is said, extract meaning in real time, and formulate and produce relevant, intelligible responses at a conversational pace’* (de Jong et al., 2009, 43). However, after reviewing the literature, it does not appear to be looking into those abilities (see analysis).

3. *To what extent does the test task reflect the construct definition?*  
Quality more or less satisfied - The test does indeed check whether candidates can track the sentences by checking if they can reproduce them, answers questions correctly, and give correct antonyms. However, as mentioned previously, the construct basis is questionable in that the literature basis is weak (see analysis).

4. *To what extent do the scoring procedures reflect the construct definition?*  
Quality satisfied - Scoring happens on four bases: vocabulary, pronunciation, syntax and fluency. These match with the construct definition. However, the exact scoring cannot be found anywhere in either Kerkhoff, et al. (2005) or de Jong, et al. (2009).

5. *Will the scores obtained from the test help make the desired interpretations about test takers’ language ability?*  
Quality not satisfied - The construct’s basis is not entirely correct nor well based, therefore the interpretations made about the language abilities will not be very helpful in the decisions to be made.

**Possible sources of bias in the task characteristics.**

6. What characteristics of the test SETTING are likely to cause different test takers to perform differently?  
Some - Noise might play a factor but every setting should be more or less equal.

7. What characteristics of the test RUBRIC are likely to cause different test takers to perform differently?  
Some - Rubric may be too easy or too difficult for some.
8. What characteristics of the test INPUT are likely to cause different test takers to perform differently?

Some - Input-length is not equally difficult for everyone.

9. What characteristics of the EXPECTED RESPONSE are likely to cause different test takers to perform differently?

Some - Length or amount of recall is not the same for everyone. However, this would make sense.

10. What characteristics of the RELATIONSHIP BETWEEN INPUT AND RESPONSE are likely to cause different test takers to perform differently?

Some - The amount of recall in repetition is likely to differ.

**Practicality.**

1. What type and relative amounts of resources are required for: (a) the design stage, (b) the operationalization stage, and (c) the administration stage?

Fully satisfied - All stages involve equipment and software from Ordinate and telephones on embassies as well as other locations.

2. What resources will be available for carrying out (a), (b), and (c) above?

Fully satisfied - All resources have been paid for by the Dutch Government.

**Interactivenss.**

*Involvement of the test takers' topical knowledge.*

1. To what extent does the task presuppose the appropriate area or level of topical knowledge, and to what extent can we expect the test takers to have this area or level of topical knowledge?

More or less satisfied - They state that every question can possibly be answered by a twelve-year old from no specific origin. However, this is not tested. Moreover, the speed of speech is relatively fast (indicated by several test takers).
2. **To what extent are the personal characteristics of the test takers included in the design statement?**

Not satisfied - Personal characteristics (besides them being learners of Dutch) have not been incorporated into the design.

3. **To what extent are the characteristics of the test tasks suitable for test takers with the specified personal characteristics?**

Characteristics have not been defined.

**Involvement of the test takers' language knowledge.**

4. **Does the processing required in the test task involve a very narrow range or a wide range of areas of language knowledge?**

Narrow range - The test only focuses on listening and speaking as measured by vocabulary, pronunciation, syntax and fluency. They cannot show their free speech.

**Involvement of language functions in the test tasks.**

5. **What language functions, other than the simple demonstration of language ability, are involved in processing the input and formulating a response?**

Answering questions, repeating sentences and giving antonyms.

6. **To what extent are the test tasks interdependent?**

Not interdependent - There is no need to have the first sentence correct in order to have the second one correct.

7. **How much opportunity for strategy involvement is provided?**

Not much - Answers are set. Strategies in how to answer part of the repetition sentences correctly can be found. However, the question remains whether this strategy can be considered positive. As participants indicated they did not necessarily understand everything but knew how to pass the test, this strategy involvement can be considered negative.
8. *Is this test task likely to evoke an affective response that would make it relatively easy or difficult for the test takers to perform at their best?*

Yes - Test takers claim that they do not understand why they have to do the test this way, are very nervous for it and does not really test their ability of Dutch (information gathered from news-programs, LinkedIn and interviews with test takers).

**Authenticity.**

1. *To what extent does the description of tasks in the TLU domain include information about the setting, input, expected response, and relationship between input and response?*

More or less satisfied - Although the exact items have not been specified, the setting, input and expected response are clear. The test is administered via telephone and test takers listen to the input divided over three tasks which they either have to repeat, oppose or answer. The relationship between the input and the response is not very clear. See below.

2. *To what extent do the characteristics of the test task correspond to those of TLU tasks?*

Not satisfied – Basically, this entails the correspondence between the items in the test with actual language use (Chapelle, 2001; Bachman & Palmer, 1996). Even though the test developers claim that it is normal to repeat sentences this cannot be found in the literature referred to (van Baaren et al., 2003). Antonyms may be useful in actual language use but are unlikely to be found in everyday speech. People do not respond to each other in opposite words or sentences. The final subtest, short questions, might be present in everyday speech, although questions with only two options are always asked. It should be noted that teachers who were interviewed for this thesis claim that these questions are too easy, especially for higher educated people. Also, they feel the correspondence between actual language use and the test is low.
Impact.

Impact on test takers.

1. To what extent might the experience of taking the test or the feedback received affect characteristics of test takers that relate to language use (such as topical knowledge, perception of the target language use situation, areas of language knowledge and use of strategies)?

   Hardly - The test is high stakes – interviews with test takers show that all of them were very nervous because so much depended on passing the test. The test is hardly authentic: the abilities needed in the test are not abilities needed in everyday life. This made it difficult for the test takers to see the use of the test. Moreover, they felt they had to study hard for it in order to pass the test. Some reported that they felt the need to use a trick to pass the test even though they thought their Dutch was good enough.

2. What provisions are there for involving test takers directly, or for collecting and utilizing feedback from test takers in the design and development of the test?

   None - Test takers have not been asked any feedback in designing the test.

3. How relevant, complete, and meaningful is the feedback that is provided to test takers?

   Not satisfied - Although most test takers know the subparts of the test beforehand, recent discussions show that test takers do not know what the computer bases the score on exactly. They would like to know what aspects they did or did not do correctly. Currently (July 2013) the government is working on more transparency.

4. Are decision procedures and criteria applied uniformly to all groups of test takers?

   More or less - The automated speech system judges every test taker equally (although the TV-program Kassa showed that especially people from Asia and Spanish-speaking countries were negatively judged by the automated system. However, this claim could not be researched within the scope of this thesis).

5. How relevant and appropriate are the test scores to the decisions to be made?

   Unknown - The literature review in the analysis above shows that the test has a weak theoretical base. The scores, and thus decisions made accordingly, are therefore questionable.
6. **Are test takers fully informed about the procedures and criteria that will be used in making decisions?**

No - Test takers more or less know what the test assesses. However, how the test is scored exactly is unclear.

7. **Are these procedures and criteria actually followed in making the decisions?**

Yes - Although this cannot be checked within the scope of this thesis, it appears that the procedures and criteria set are actually followed.

*Impact on teachers.*

8. **How consistent are the areas of language ability to be measured with those that are included in teaching materials?**

Consistent - The teaching-materials have been adjusted to match the test. They now involve training for all subtests on the TGN (and other exams). Teachers who use other ‘regular’ books which focus on the complete learning of Dutch still use the adjusted materials as well to be able to prepare their students (i.e. negative washback).

9. **How consistent are the characteristics of the test and test tasks with the characteristics of teaching and learning activities?**

Not consistent - Teachers state that they do not want to teach to the test but do want their pupils to pass the test. Therefore, teachers tailor their activities to the test as teachers feel this is the only way their pupils can pass.

10. **How consistent is the purpose of the test with the values and goals of teachers and of the instructional program?**

Not consistent - Interviews with teachers show that they agree that some sort of measurement is needed to evaluate the level of Dutch of these people. The measurement used does not match with that what teachers want to learn their students. They want to invest their time and energy in the learning of Dutch not in training for the exam.

*Impact on society and education systems.*

11. **Are the interpretations we make of the test scores consistent with the values and goals of society and the education system?**

No - The goal of the test is to enforce people to learn Dutch in order to be part of the society. The news-programs in the beginning of 2013 have shown that people who pass the tests do
not necessarily speak Dutch. Therefore, the interpretations are not necessarily consistent with ideas in society.

12. To what extent do the values and goals of the test developer coincide with those of society and the education system?

Hardly - Both teachers and test takers do not understand why this test is in the Civic Integration Exam. All teachers and test takers agree that a test is needed but believe that the TGN is not the best test for its purpose.

13. What are the potential consequences, both positive and negative, for society and the education system, of using the test in this particular way?

Test takers learn to the test: they try to pass the test by practicing the subtests or use tricks but end up speaking Dutch below the A2 level or do not speak Dutch at all.

14. What is the most desirable positive consequence, of the best thing that could happen as a result of using the test in this particular way, and how likely is this to happen?

The most desirable positive consequence would be that test takers have actually learned Dutch on the A2 level. This is likely to happen but does not necessarily happen all the time.

Washback

The interviews with the test takers and teachers show that washback can be found in both teaching and learning. Although teachers do not want to teach to the test and may choose not to at first, they start teaching to the test by the end of the course as the test is too difficult for the immigrants to do unprepared. They give practice-tests and lists of antonyms to learn by heart. Test takers consequently learn to the test (even when they did not take a course) as their goal is to become Dutch, not necessarily to learn Dutch. Test takers even claim to not be able to speak Dutch when they passed the test.

Panel research

Even though care was taken to judge this test as objectively as possible, this evaluation is naturally subjective (Bachman & Palmer, 1996). Therefore, a panel of experts in the field of Dutch as an L2 and/or testing was set up to evaluate this analysis. Unfortunately, one member could not finish the evaluation. It should be noted that this expert was very interested in the subject of this thesis and was very willing to evaluate the analysis. Therefore, the results from the two other experts are reported below.
Overall the agreement of the two experts with the previous evaluation of usefulness is substantial. Before the panel could start the actual evaluation by giving an overall judgement per result section. Here, some interesting ideas were noted. Panel member 1 (PM1) pointed out some issues with the TGN which were not mentioned above: Due to the high consequences of the test, a test which measures exact levels is needed. The TGN as it is today is too global measurement and therefore not good enough. Moreover, PM1 notes that the high level of reliability of the test is likely due to the use of differing levels of participants’ proficiency. Another problem PM1 has with the TGN is the automated speech recognition system: the speech system first needs to recognise the utterance as a certain utterance before it can score it, if the word is correct but mispronounced the utterance is scored as wrong. What is tested, then, is pronunciation. The second panel member (PM2) points out that when a person is limited in language proficiency the first subtest will only depend on imitation. Moreover, PM2 states that money was too big an issue in the choice of the automated system: the Ministry of Justice wanted a good test but did not wish to pay for it which led to a test which is far from ideal.

The more in-depth questions per quality were judged as well. The descriptions of the qualities practicality, interactiveness, and authenticity were approved by the panel. Therefore, only the other qualities are discussed below.

The quality reliability received some additions: PM1 wonders whether all test takers are familiar with the testing method and whether the instructions are understood by everyone. Moreover, the random selection of items can lead to differing test-versions which can be more or less difficult thus more or less reliable. Finally, PM1 questions whether words with two possible opposites (such as arm – rijk/been) are scored as correct. PM2 doubts if the rubric motivates why the subtests repetition, opposites, and short answers are tested.

Construct validity already drew considerable negative criticism in the evaluation mentioned above but both panel-members add to this that the score ‘quality more or less satisfied’ for the ‘clear and unambiguously definition’ (question one) of the construct is rather mild as the construct did not prove to be solid. Moreover, PM2 states that the test leans heavily on imitation which was not its original goal. Both members also mention that, although the basis of the score is clear, the actual scoring by the system is unknown. Both panel-members also believe that the test may very likely be biased towards people from Asia or towards other Mother Tongues.

Finally, the panel-members add that the impact of the test is substantial because the test and actual language are not in alignment which leads to the use of materials to prepare for the test. PM2 adds that a good test leads to ‘good’ behaviour in class or in preparation for the test. Ideally then, the test should be based on communication.
The next chapter discusses the results for the evaluation of usefulness, washback, and the panel-review mentioned above.

Discussion

The goal of this research was to get insight into the validity of the TGN by use of the model for the evaluation of usefulness (Bachman & Palmer, 1996), the test’s washback on teaching & learning and the panel-review. The results for these three methods are discussed respectively next.

Evaluation of usefulness

The discussion of the results of the six qualities is done according to the Bachman & Palmer (1996) model for the evaluation of usefulness. This model is prescriptive in that it requires language tests to be authentic, i.e. corresponding to actual language use. Testing should be based in communication and in line with teaching. The results are discussed in light of this model and may not be in line with other (non-communicative or non-authentic) approaches.

Reliability. The first quality refers to the consistency of measurement of the TGN. Research done by de Jong et al. (2009) and Kerkhoff, et al. (2005) indicate an overall high reliable score of .94. The automated speech analysis system was also compared with judgements of people who were trained to rate in the same way the system does: scores between .80 and .94 were found. However, this training is likely the reason for the high agreement between the two raters. Perhaps, human-raters would not have given the same scores when they were not trained.

The analysis according to the model used did show some reliability issues: there may not be equal amounts of noise on the different locations and because the items are randomly chosen from a database the reliability scores may not be the same in every test. Unfortunately, it was impossible to complement the results from earlier work as the official test was not made available. We should therefore rely on the scores given in the earlier work.

Although the TGN is reliable, there still can be validity issues with the test. Several researchers have pointed out the importance of the other qualities (Chun 2009; Fulcher, 2010; Yang et al., 2002; Xi, 2010) and the insufficiency of comparisons of automated systems’ judgments with human-raters (Bejar & Bennett, 1997). Therefore, the construct validity is discussed next.
**Construct validity.** The test’s justification was reviewed in order to get insight into the construct validity of the test because this document (Kerkhoff *et al.*, 2009) should provide the reasons why this type of test was chosen. The review indicated which language abilities should be tested: understanding sentences and replying to them. However, it is unclear why the subtests repetition, short answers to questions, and antonyms were chosen. The latter two subtests are not based on any literature whatsoever. For the first subtest some theoretical basis is given in referring to research by Baddely (1986;2000) who claims that people can remember sentences up to twenty words when the words are meaningfully connected. People fail to do this when the connection is not present after seven words. However, the practice test shows that only 36% of the sentences contain more than seven words which would mean that the other 64% would be mere imitation. Considering that the test-developers claim the practice-test to be a representative example of the actual TGN this would mean that over half of the items in the TGN are possibly not useful for the purposes of the test. It should again be noted that the actual TGN could not be used for the current purposes. These results thus cannot be checked for the real test which may perhaps yield different results.

**Practicality.** This quality is satisfied. When CINOP proposed the TGN together with Language Testing Services and Ordinate, a budget was set up and agreed upon by the Ministry. The Ministry has made all resources available. Note that this quality should be looked into before the test is put to use (Bachman & Palmer, 1996).

**Interactiveness.** The interactiveness of the TGN is rather low. Test takers do not need to involve any topical knowledge as the test-developers claim that a twelve year old with no specific knowledge of the Netherlands should be able to make the test. However, this claim has not been researched so it is questionable whether this is true. Furthermore, the test does not include personal characteristics and involves a narrow range of areas of language knowledge. The items are not interdependent. And most importantly, the test is not appreciated by the test takers which affects their nerves and makes them use tricks to pass the test. Again, this test quality is considered important by Bachman & Palmer but may not be important in other models.

**Authenticity.** According to Bachman & Palmer (1996) and Chapelle (2001) a (computerised) test should have a high degree of authenticity. The current results show that the analogy between the TGN and the target language use is rather low. The same was found by Chun (2009) who used the Bachman & Palmer (1996) model for the PhonePass test. The current research found differing degrees of correspondence per subtest. The first subtest (repetition) can hardly be considered authentic. That is, in normal conversation we do
not parrot each other. The test-developers do attempt to base their choice for repetition on Van Baaren et al. (2003) but this research merely claims that when waitresses repeat orders from customers bigger tips are given. Moreover, they do not claim that repetition is normal in everyday communication. The second subtest (antonyms) requires vocabulary knowledge which may be considered more useful in everyday language use. It remains questionable whether people actually use antonyms outside language tests on an everyday basis as this could not be researched within this thesis. The third subtest requires the answering of questions which, from a communicative view of language, can be considered more authentic. We use questions on a daily basis although these may not always be two-fold like the questions in the TGN. Therefore, using the Bachman & Palmer (1996) model only the latter subtest can be considered authentic.

**Impact.** The impact on both test takers and teachers is substantial. The first group reports nerves, anxiety and fear of making the test as the score on the test decides whether or not they can come or stay in the Netherlands. Embse & Hasson (2012) also reported high levels of anxiety and stress in high-stakes test situations. Moreover, test takers report that they would have liked to know what criteria were used for their passing or failing on test. Fortunately, since July 2013 test takers can gain more insight in their scores on the four aspects (vocabulary, syntax, pronunciation, and fluency) of the TGN. However, it should be noted that only four people who are all university students have been interviewed. Other results could perhaps have been found when a more diverse and larger group of test takers were interviewed.

Teachers report issues with the TGN as well. Their idea of a good test is not in alignment with the TGN which forces most of them to alter their curriculum to the test (i.e. washback, which will be discussed in more depth below) whilst they feel that the preparation for the test is a waste of time: people do not learn Dutch by learning to this test. Again, these results are limited in that only six teachers were interviewed for the current purposes.

The impact on society, although not explicitly researched in this thesis, can be substantial as well. The introduction of this thesis reported TV-programs which showed people who passed the test but did not speak Dutch and people who do speak Dutch but failed the test. Both outcomes are undesirable.

The evaluation of all six test qualities thus showed that the construct validity, interactiveness, authenticity, and impact of the TGN are insufficient. This is likely due to the way the language ability is defined and construed in the test. According to Bachman & Palmer (1996) an ill-defined construct can lead to a lower construct validity as well as lower degrees on authenticity and interactiveness. Moreover, the lower the authenticity of a test is,
the more impact the test is likely to have on stakeholders. Becker & Ribeiro (2005) researched the Phonepass test and found that the construct was not in line with linguists’ definition of ‘fluency’. Xi (2010) also pointed out that the speaking ability construct of the Phonepass lacked in representation which, in turn, leads test takers to use tricks to pass the test and to adjust their learning to the test. It is thus highly likely that the construct of the TGN caused lower degrees of interactiveness and authenticity, which in turn caused the higher impact on stakeholders which can result in negative washback. Whether this is true for the TGN is discussed next.

Washback

In the Bachman & Palmer (1996) model only two questions are specifically devoted to the phenomenon washback in teaching. The theoretical framework presented earlier on washback showed considerable effects of high-stakes tests on teaching and learning (Alderson & Wall, 1993; Bailey, 1996; Cheng, et al., 2004). Therefore, special attention was paid to the phenomenon by conducting interviews with both teachers and test takers as previous research advocated for more research into washback (Cheng & DeLuca, 2011; Girard & Haniford, 2006).

The interviews showed that teachers tailor part of their classes to the TGN. Although most teachers agree that the goal should be the learning of Dutch, teachers will teach to the test in the end because the stakes of the test are high. Moreover, they want their pupils to pass the test. Both make it impossible for them to not teach to the test by the end of a course. They either buy or produce materials which are tailored to the test.

Test takers learned to the test. They used the occasional trick and learned the antonyms by heart. They indicate that, although these ‘tricks’ made them pass the test, their Dutch did not improve. Therefore, most pursued the higher levelled State Exam I (STEX I) as it made them use the language more frequently. It should be noted that only higher-educated test takers (with obvious higher study abilities) were interviewed in this research. Interviews with illiterate or lower-educated test takers may have yielded other results.

Both the interviews with teachers and test takers showed considerable washback. This washback can be considered negative: the test is not in alignment with the goals of teachers and test takers, both groups of interviewees claim that they do not want to prepare for the test this way, and people do not necessarily learn Dutch. These results align with Burrows’ (2004) research into Australia’s Adult Migrant English Program in which clear washback effects were found although teachers could chose not to teach to the test.
The previous section indicated that the high level of impact may have an affect on the washback of the TGN. This appears to have happened: both teachers and test takers refer to the high stakes of the test and their adjustment of teaching or learning to the test. However, different results may be found in lower-educated, analphabetic test takers or a higher number of test takers (and teachers).

Panel-review

The evaluation of usefulness was sent to a panel of three experts in the field of testing to limit the possible subjectivity of the current research. Only two were able to finish the review. The overall agreement with the evaluation of the experts turned out to be substantial: complete agreement was found in the qualities practicality, interactivity, and authenticity. Whilst for the first this makes sense (the test has already been paid for), this does not mean the latter two should be approved of. This approval is likely due to their belief that language testing should be more authentic (especially in this high-stakes test) and their disapproval of the TGN as a measurement of spoken language ability. The latter can also be found in their additions to the other qualities: the panel added some doubts to how reliable the test really is as it is a very global measurement and whether the random selection of items affected the reliability. Moreover, the panel doubts if the rubric actually motivates why the subtests are chosen. In reference to the test as a whole, one member adds that money was too big an issue in the development of this test which led to an imperfect test. Furthermore, both experts express that the construct basis is not as solid as it should be and causes substantial impact as the test and actual language use are not corresponding with each other.

The results of the panel-review thus support the evaluation of usefulness. The panel was very explicit in the need for more authentic testing which is likely the reason for their ample agreement. Although both panel-members are considered experts in the field, only two managed to finish the review. Other results, or more additions, would perhaps be yielded with a larger panel.

The next chapter will report the key findings and implications of this research.

Conclusion

The key findings are presented firstly in this chapter. These are followed by implications for the use of the TGN and implications for future research.

The aim of this thesis was to get insight into the validity of the TGN as several issues with the test were reported in the news and other media. This was done by using a model for the evaluation of usefulness by Bachman & Palmer (1996) which distinguishes six test
qualities: reliability, construct validity, practicality, interactiveness, authenticity, and impact. All qualities should be present in any given test, especially in high-stakes tests such as the TGN. This model was completed with interviews which were conducted with test takers and teachers to get insight into the effects of the TGN on teaching and learning. Finally, the evaluation was shown to a panel of experts who reviewed it and provided it with feedback.

The results show that the TGN is weak on the qualities construct validity, authenticity, interactiveness, and impact. Negative washback on both teaching and learning was found as well. The expert panel that reviewed the results agreed upon these results and added interesting considerations.

The findings suggest that the TGN is not useful for the current purposes as the only measurement of spoken language ability in the Civic Integration Exam (from January 2013 onwards). The previous discussion showed that this is likely due to the construct used in the test. Considering the PhonePass test is based on more or less the same construct, these results may also be applicable to this test, though more research should look into this.

More broadly, the methods used to get insight into the washback of the test appear to be good tools for the purposes. Moreover, the results for washback confirm earlier research: high-stakes tests are likely to affect teaching and learning. Therefore, in the future, test-developers should keep the ultimate goal in mind and develop the test accordingly. It should be noted that the model which was used to get insight into the validity is rather prescriptive in that it advocates for authentic, communicative language testing. Although the panel agreed with the model, other models for evaluating the validity of a test can yield other results.

**Future**

The assertion that the TGN is not a valid test for its purpose is one thing, finding an alternative is another. More research into alternatives or solutions for the TGN would be a step in the right direction. This research should be done by independent institutions or universities as the research for the TGN in its present form was done by its developers. In line with Ryan (2002), we argue that an independent evaluator should look into this.

If the government insists on using automatically evaluated speech systems, care should be taken in choosing the construct. In the current TGN the construct proved to be the reason why the test is not as valid as it could possibly be. Messick (1996) and Bachman & Palmer (1996) indicated that a test is likely to be considered positively when the degree of correspondence between the test situation and actual language use is highly authentic.

The interviews with teachers also showed suggestions for alternatives. Some teachers claim that the Civic Integration Exam needs a test which actually tests communicative abilities. They prefer a communicative examination with actual people.
Teachers do understand that the large amount of money which is needed for such an exam is not likely to be present. This is the reason why most teachers advocate for the return of the portfolio, in which people actually had to communicate. This portfolio may or may not be combined with the TGN.

Moreover, extensive research is needed into the washback and impact of high-stakes tests like the TGN. The current research merely touched upon the phenomena but the results indicate that these are vital in the acceptance of a test. The interviews showed a great deal of teachers’ interest in thinking about possible alternatives for the TGN. We therefore advocate for consultations between teachers, test-developers and the government to either improve the TGN or think about alternatives which satisfy all stakeholders.

Overall, future research is needed into the validity and washback of low- as well as high-stakes tests, before and when a test is put to use. Test-developers need to be on top of a test’s validity.
References


Appendices
Practice-test TGN

Dutch telephone number: +3188-7890123

Welkom bij het oefenexamen Nederlandse taal.


Nu is het uw beurt. Luister naar de zin en zeg precies na wat u hoort.

Het moet in januari klaar zijn.
Het is niet helemaal gegaan zoals we verwacht hadden.
Heb je een pen bij je?
Dat kan wel kloppen.
De volgende keer betaal ik.
Dat kan iedereen wel zeggen.
Daar heb ik nog nooit van gehoord.


Nu is het uw beurt. Luister naar de vraag en geef antwoord.

Is een auto om in te rijden of in te koken?
Kun je melk eten of drinken?
Wat is langer, een arm of een been?
Fiets je op een rivier of op een pad?
Één uur, hoeveel kwartier is dat?
Als je een groot gezin hebt, heb je dan veel of weinig kinderen?
Hoe noem je het gebouw waar kinderen les krijgen?
Wie woont er op een boerderij?

Nu is het uw beurt. Luister naar de zin en zeg precies na wat u hoort.

We gaan er de volgende les mee verder. 8
De aardappels zijn op. 4
Dat kun je op je vingers natellen. 7
Ik moet een nieuwe bril. 5
Heb je terug van vijftig? 5
Door de harde regen zijn veel planten beschadigd. 8
Twee is teveel. 3


Nu is het uw beurt. Luister naar het woord en zeg het tegengestelde woord.

Arm

Ochtend

Laatste

Oorlog

Achterin

Gister


Renate is bang voor vliegtuigen. Ze durft niet te vliegen. Ze is zelfs bang voor een vliegtuig dat over komt vliegen. Haar moeder vertelt dat Renate toen ze klein was, ook al bang was voor vliegtuigen. Als er een vliegtuig over kwam vliegen, dook ze altijd onder de tafel of onder een stoel.

*Dank u voor het bellen. U kunt nu ophangen.*
Interview questions

Teachers
1. Who are you and what do you do?
2. How would you describe the organisation you work for?
3. How long have you been teaching?
4. What does a ‘standard’ class look like?
5. Is that normal in comparison with the other teachers?
6. How do you feel about the changes in the new Civic Integration Exam?
7. In the new exam, is there something you miss?
8. How do you generally prepare for KNS?
9. How do you generally prepare for TGN?
10. What materials do you use for TGN?
11. What troubles test takers the most?
12. Do you feel that you should spend more time on certain aspects that are not in the exam?
13. How do you feel about the recent discussion about the TGN?
14. (optional) Do you have to teach tricks to your students?

Test takers
1. Who are you and what do you do?
2. When did you take the Civic Integration Exam?
3. Why did you have to do it?
4. Did you take a course?
5. What did it look like?
6. Did you practice a lot?
7. What kinds of materials were used in class? Which materials did you use?
8. How did you generally prepare for the exam?
9. And more specifically, how did you prepare for the sub-parts?
10. What do you think about those parts?
11. How did you practice for the TGN?
12. How did you practice for KNS?
13. How did you feel about the TGN? And KNS?
14. Were the tests difficult?
15. How much time in class was spent on TGN? How?
16. How useful was the Civic Integration Exam for you?
17. (optional) Do you think you could have learned more Dutch if you did not have to practice for the TGN?
Panel-research for the evaluation of usefulness TGN

Name:
Date:

Geacht panellid,

Met behulp van dit formulier kunt u aangeven in hoeverre u het eens bent met de evaluatie die gemaakt is van de TGN. Allereerst volgt de mogelijkheid tot reageren op de verschillende onderdelen die aan het begin staan van de analyse, welke u kunt vinden in het andere document. Dit is niet verplicht, maar biedt u de mogelijkheid tot reageren op de informatieverzameling.


Hartelijk dank voor uw medewerking,

Anne Abeling
Overall remark on ‘Why the TGN?’

Overall remark on ‘Administration & construct TGN’

Overall remark on ‘Literature review of justification TGN’

Overall remark on ‘Comparison of TGN with CEF’

Overall remark on ‘Interviews stakeholders’

Overall remark on ‘Evaluation of the usefulness of the TGN’
<table>
<thead>
<tr>
<th>Reliability</th>
<th>Extent to which quality is satisfied</th>
<th>Explanation</th>
<th>Extent of agreement</th>
<th>Explanation (facultative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. To what extent do characteristics of the test setting vary from one administration of the test to another?</td>
<td>Quality completely satisfied</td>
<td>There is no variation in setting. The test is administered via telephone. The noise-levels may differ on locations.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>2. To what extent do characteristics of the test rubric vary in an unmotivated way from one part of the test to another, or on different forms of the test?</td>
<td>Quality completely satisfied</td>
<td>Instructions or rubrics are the same for everyone.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>3. To what extent do characteristics of the test input vary in an unmotivated way from one part of the test to another, from one task to another, and on different forms of the test?</td>
<td>Quality satisfied</td>
<td>The test's input is the same for everyone at any time even though items are different. This cannot be avoided.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>4. To what extent do characteristics of the expected response vary in an unmotivated way from one part of the test to another, or on different forms of the test?</td>
<td>Quality completely satisfied</td>
<td>The expected (i.e. correct) answers do not change and have not changed in the past years.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>5. To what extent do characteristics of the relationship between input and response vary in an unmotivated way from one part of the test to another, or on different forms of the test?</td>
<td>Quality completely satisfied</td>
<td>The test is the same for everyone at any time. The difference in parts of the test have been motivated. This (and other parts of the test's reliability) has been evaluated deeply (see</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>Construct validity</td>
<td>Extent to which quality is satisfied</td>
<td>Explanation</td>
<td>Extent of agreement</td>
<td>Explanation (facultative)</td>
</tr>
<tr>
<td>--------------------</td>
<td>-------------------------------------</td>
<td>-------------</td>
<td>---------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>18. Is the language ability construct for this test clearly and unambiguously defined?</td>
<td>Quality more or less satisfied</td>
<td>The language ability has been explained clearly: it seeks to see how fast a candidate can understand the meaning and can reply to it (Kerkhoff, et al., 2005). However, why the different subtests are chosen is not clear at all. This should be added and explained in depth.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>19. Is the language ability construct for the test relevant to the purpose of the test?</td>
<td>Quality not satisfied</td>
<td>The test’s purpose was to get insight into the listening and speaking ability (as asked by the Dutch Ministry of Justice) as measured by ‘the facility with which candidates are able to track what is said, extract meaning in real time, and formulate and produce relevant, intelligible responses at a conversational pace’ (de Jong et al., 2009, 43). However, after reviewing the literature, it does not appear to be looking into those abilities (see analysis).</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>20. To what extent does the test task reflect the construct definition?</td>
<td>Quality more or less satisfied</td>
<td>The test does indeed check whether or not candidates can follow the sentences by checking if they can</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
</tbody>
</table>
reproduce them, answers questions correctly, and give correct antonyms. However, as mentioned previously, the construct basis is questionable in that the literature basis is weak and sometimes poorly defined or extracted (see analysis).

<table>
<thead>
<tr>
<th>21. To what extent do the scoring procedures reflect the construct definition?</th>
<th>Quality completely satisfied</th>
<th>Scoring happens on four bases: vocabulary, pronunciation, syntax and fluency. These match with the construct definition.</th>
<th>Agree/Disagree</th>
</tr>
</thead>
</table>

**Possible sources of bias in the task characteristics**

<table>
<thead>
<tr>
<th>23. What characteristics of the test SETTING are likely to cause different test takers to perform differently?</th>
<th>Hardly any</th>
<th>Noise might play a factor but in principle every setting is more or less equal.</th>
<th>Agree/Disagree</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>24. What characteristics of the test RUBRIC are likely to cause different test takers to perform differently?</th>
<th>Hardly any</th>
<th>Rubric may be too easy or too difficult for some.</th>
<th>Agree/Disagree</th>
</tr>
</thead>
</table>

<p>| 25. What characteristics of the test INPUT are likely to cause different | Hardly any | Input-length is not equally difficult for everyone. | Agree/Disagree |</p>
<table>
<thead>
<tr>
<th>Test Takers to Perform Differently?</th>
<th>Hardly Any</th>
<th>Length or Amount of Recall is Not the Same for Everyone. However, this Would Make Sense.</th>
<th>Agree/Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>26.</strong> What characteristics of the Expected Response are likely to cause different test takers to perform differently?</td>
<td>Some</td>
<td>The amount of recall in repetition is likely to differ.</td>
<td>Agree/Disagree</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Extent to which Quality is Satisfied</th>
<th>Explanation</th>
<th>Extent of Agreement</th>
<th>Explanation (facultative)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>16.</strong> To what extent does the description of tasks in the TLU domain include information about the setting, input, expected response, and relationship between input and response?</td>
<td>More or less satisfied</td>
<td>Although the exact items have not been specified, the setting, input and expected response are clear. The test is administered via telephone and test takers listen to the input divided over three tasks which they either have to repeat, oppose or answer. What the relationship between the input and the response is, is not very clear. See below.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td><strong>17.</strong> To what extent do the characteristics of the test task</td>
<td>Not satisfied</td>
<td>Basically this entails the correspondence between the items in the test with actual language</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
</tbody>
</table>
correspond to those of TLU tasks? Even though the test developers claim that it is normal to repeat sentences this cannot be found in the literature referred to (van Baaren et al., 2003). The subtest ‘antonyms’ might be useful in actual language use but is not found in everyday speech. People do not respond to each other in opposite words or sentences. The final subtest, short questions, might be present in everyday speech. It should be noted that teachers who were interviewed for this thesis claim that these questions are too easy. Especially for higher educated people. Also, they feel the correspondence between actual language use and the test is low.
<table>
<thead>
<tr>
<th>Interactiveness</th>
<th>Extent to which quality is satisfied</th>
<th>Explanation</th>
<th>Agreement</th>
<th>Explanation (facultative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Involvement of the test takers’ topical knowledge</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18. To what extent does the task presuppose the appropriate area or level of topical knowledge, and to what extent can we expect the test takers to have this area or level of topical knowledge?</td>
<td>More or less satisfied</td>
<td>They state that every question can possibly be answered by a twelve-year-old from no specific origin. However, this is not tested. Moreover, the speed of speech is relatively fast (indicated by several test takers). Such a speed is not necessarily asked for the level of A2.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>19. To what extent are the personal characteristics of the test takers included in the design statement?</td>
<td>Not satisfied</td>
<td>Personal characteristics (besides them being learners of Dutch) have not been included.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>20. To what extent are the characteristics of the test tasks suitable for test takers with the specified personal characteristics?</td>
<td>Not satisfied</td>
<td>Characteristics have not been defined.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>Involvement of the test takers’ language knowledge</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21. Does the processing required in the test task involve a very narrow range or a wide range of areas of language knowledge?</td>
<td>Narrow range</td>
<td>The test only focuses on listening and speaking as measured by vocabulary, pronunciation, syntax and fluency.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>Involvement of language functions in the test tasks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22. What</td>
<td>Answering</td>
<td></td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>Question</td>
<td>Response</td>
<td>Justification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>----------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>To what extent are the test tasks interdependent?</td>
<td>Not</td>
<td>There is no need to have the first sentence correct in order to have the second one correct.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How much opportunity for strategy involvement is provided?</td>
<td>Not very high</td>
<td>Answers are set. Strategies in how to answer part of the repetition sentences correctly can be found.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is this test task likely to evoke an affective response that would make it relatively easy or difficult for the test takers to perform at their best?</td>
<td>No</td>
<td>Test takers claim that they do not understand why they have to do the test this way, are very nervous for it and does not really test their ability of Dutch (information gathered from news-programs, LinkedIn and interviews with test takers).</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Impact on Test Takers

<table>
<thead>
<tr>
<th>Impact</th>
<th>Extent to which quality is satisfied</th>
<th>Explanation</th>
<th>Agreement</th>
<th>Explanation (facultative)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>26. To what extent might the experience of taking the test or the feedback received affect characteristics of test takers that relate to language use (such as topical knowledge, perception of the target language use situation, areas of language knowledge and use of strategies)?</strong></td>
<td>Hardly</td>
<td>The test is high stakes – interviews with test takers show that all of them were very nervous because so much depended on passing the test. The test is hardly authentic: the abilities needed in the test are not abilities needed in everyday life. This made it difficult for the test takers to see the use of the test and they felt they had to study hard for it in order to pass the test. Some reported to feel the need to use a trick to pass the test even though they felt their Dutch was good enough.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td><strong>27. What provisions are there for involving test takers directly, or for collecting and utilizing feedback from test takers in the design and development of the test?</strong></td>
<td>None</td>
<td>Test takers have not been asked any feedback in designing the test.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td><strong>28. How relevant, complete, and meaningful is the feedback that is provided to test takers?</strong></td>
<td>Not satisfied</td>
<td>Although test takers know the subparts of the test, recent discussions show that test takers do not know what the computer bases the score on exactly. They would like to know exactly what aspects they did or did not do correctly.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>Question</td>
<td>Answer</td>
<td>Explanation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>--------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29. Are decision procedures and criteria applied uniformly to all groups of test takers?</td>
<td>Yes</td>
<td>The automated speech system judges every test taker equally (although the TV-program <em>Kassa</em> showed that especially people from Asia and Spanish-speaking countries were negatively judges by the automated system. However, this claim could not be researched within the scope of this thesis).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30. How relevant and appropriate are the test scores to the decisions to be made?</td>
<td>Hardly</td>
<td>The literature review in the analysis above shows that the test is ill based. The scores and thus decisions made are therefore questionable.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31. Are test takers fully informed about the procedures and criteria that will be used in making decisions?</td>
<td>No</td>
<td>Test takers more or less know what the test assesses. However, how the test is scored exactly is unclear.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32. Are these procedures and criteria actually followed in making the decisions?</td>
<td>Yes</td>
<td>Although this cannot be checked it seems that the procedures and criteria set are actually followed.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Impact on teachers**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>33. How consistent are the areas of language ability to be measured with those that are included in teaching materials?</td>
<td>Consistent</td>
<td>The teaching-materials have been adjusted to the test. They now involve training for all subtests on the TGN (and other exams). Teachers who use other books which focus on the complete learning of Dutch still use the adjusted materials as well to be able to prepare their students (i.e. washback).</td>
</tr>
<tr>
<td>34. How</td>
<td>Not</td>
<td>Teachers state that they</td>
</tr>
</tbody>
</table>

Agree/Disagree
### Impact on society and education systems

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>35. How consistent is the purpose of the test with the values and goals of teachers and of the instructional program?</td>
<td>Not consistent</td>
<td>Interviews with teachers show that they agree that some sort of measurement is needed to evaluate the level of Dutch of these people. The measurement used does not match with that what teachers want to learn their students. They want to invest in the learning of Dutch not in training for the exam.</td>
</tr>
<tr>
<td>36. Are the interpretations we make of the test scores consistent with the values and goals of society and the education system?</td>
<td>No</td>
<td>The goal of the test is to enforce people to learn Dutch in order to be part of the society. The news- programs in beginning of 2013 have shown that people who pass the tests do not necessarily speak Dutch. Therefore, the interpretations are not necessarily consistent with ideas in society.</td>
</tr>
<tr>
<td>37. To what extent do the values and goals of the test developer coincide with those of society and the education system?</td>
<td>Hardly</td>
<td>Both teachers and test takers do not understand why this test is in the ‘inburgeringsexamen’. All teachers and test takers agree that a test is needed but</td>
</tr>
<tr>
<td>38. What are the potential consequences, both positive</td>
<td>-</td>
<td>Test takers learn to the test, that is, they try to pass the test by practicing the subtests or</td>
</tr>
</tbody>
</table>
and negative, for society and the education system, of using the test in this particular way?

<table>
<thead>
<tr>
<th>39. What is the most desirable positive consequence, of the best thing that could happen as a result of using the test in this particular way, and how likely is this to happen?</th>
</tr>
</thead>
<tbody>
<tr>
<td>The most desirable positive consequence would be that test takers have actually learned Dutch on the level of A2. This is likely to happen but does not necessarily happen all the time.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Practicality</th>
<th>Extent to which quality is satisfied</th>
<th>Explanation</th>
<th>Agreement</th>
<th>Explanation (facultative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40. What type and relative amounts of resources are required for: (a) the design stage, (b) the operationalisation stage, and (c) the administration stage?</td>
<td>Fully satisfied</td>
<td>All stages involve equipment and software from Ordinate and telephones on embassies as well as other locations.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
<tr>
<td>41. What resources will be available for carrying out (a), (b), and (c) above?</td>
<td>Fully satisfied</td>
<td>All resources have been paid for by the Dutch Government.</td>
<td>Agree/Disagree</td>
<td></td>
</tr>
</tbody>
</table>

Hartelijk dank voor uw medewerking!