Development of a Test of Spoken Dutch for Prospective Immigrants

John H.A.L De Jong, Matthew Lennig, Anne Kerkhoff and Petra Poelmans

With increasing facility of world-wide mobility many western countries are confronted with growing numbers of immigrants and are seeking means to restrain immigration. To that effect many countries are setting new or more rigorous language requirements. The Netherlands is one such country. This paper reports on the definition, development and validation of a set of automatic assessment instruments measuring a minimal level of spoken Dutch and basic aspects related to culture and life in the Netherlands. During field testing 1,300 items were submitted to two target groups: 821 native speakers and 1,522 non-native speakers from 121 different nationalities. Arguments for test validity are based on evaluation of construct-relevant and construct-irrelevant factors. The political context is discussed.

Based on a broadly supported parliamentary decision, the Ministry of Justice of the Netherlands in December 2003 commissioned the development of an examination system for testing the Dutch oral language skills of foreigners who want to immigrate permanently to the Netherlands for economic or family reasons. This assessment should take place in the country of origin prior to being admitted to the Netherlands. Moreover, the test should also be appropriate for use within the framework of an examination scheme for naturalization within the Netherlands, which would require a higher level of ability than at first entry. The reporting scale for the test should therefore represent reliable measurement at different points on the underlying ability continuum. It was further required by the Ministry of Justice that the reporting scale be related to the scales of the Common European Framework of Reference for Languages (henceforth: CEF; Council of Europe, 2001). Furthermore, the test of the Dutch language was to be accompanied by a test assessing knowledge of Dutch society, its political structure, its rules and norms and some historical and geographical facts[1]. This paper reports on some aspects of the production and validation of the language test, i.e., the test of spoken Dutch, henceforth referred to as the TGN (acronym of the Dutch name: Toets Gesproken Nederlands). For the full report in Dutch see Kerkhoff, Poelmans, De Jong, & Lennig (2005). For the development of the test assessing knowledge of the Dutch Society, see De Jong and Tijssen (2005). Both tests were developed by a consortium with main contractor CINOP (NL), and subcontractors Language Testing Services (NL) and Ordinate Corporation (USA).

Background

The political decision to put out a call for tender to develop an instrument to assess Dutch language ability and knowledge about Dutch society resulted from discussions in politics as well as in the research community about the negative consequences of the social segregation of large numbers of immigrants in the Netherlands (Odé, 2002). Lack of integration leads to a vicious circle, where parents cannot help their children, children drop

[1] See the appendix for some details on how the specifications for the test of knowledge of Dutch society were defined to take into account the minimal level of ability in the Dutch language of the candidates.

out of schools, leading to feelings of hopelessness and despair, which develop from the realization of the improbability of achieving success in terms of the values and goals of the larger society (Odé, 2002:14). In the larger cities in the Netherlands over 50% of the school generation is of non-Dutch origin, leading to the development of so called "black schools", where children underperfom in comparison to the their Dutch peers. Van de Laan (2007) points out that ethnic concentration is assumed to have negative effects on the integration of ethnic minorities, the most important cause being the lack of contact with native Dutch.

In addition, modern society has brought about a new form of slave trade. Of the 20 thousand new arrivals in the Netherlands in 2001 three quarters were women, mostly in the age range of 16 to 26 (CBS, 2003). This high number of young women can be broken down into three main streams. A first stream is composed of women from Eastern and Middle Europe (mainly Bulgaria Romania, and Russia and form Asia (mainly Thailand and the Philippines) who come to marry mostly elderly Dutch men. Contacts are established via websites leaving no doubt about their purpose with names such as www.mailorderwomen.com [2] A second stream is formed by young women from Morocco and Turkey, mainly from rural areas. Young, second generation men living in the Netherlands find the young women of their age in the Netherlands to be too westernized, lacking the traditional obedience they have seen their mothers pay to their fathers. These men depart on holiday trips to their ancestral countries to find brides who they expect to conform more to the traditional role model they have been accustomed to in their youth. The third, saddest, group is formed by young women from Eastern Europe (again) and from Africa, who are lured to a better life in the Netherlands only to land in forced prostitution. On the basis of the data out of the police investigations for the period 1997-2000, Van Dijck (2002) estimates the yearly amount of new victims of trafficking in women at approximately 3500 (see also Vocks and Nijboer 2000; Aronowitz, 2001; Hopkins and Nijboer, 2001).

The situation of segregation and the growing number of studies revealing the undesirable developments in immigration mentioned above led to a motion in Parliament (Rouvoet 1998-1999), adopted by all political parties from left to right, requiring the government to take measures. A law on integration for new residents (Wet op de Inburgering voor Nieuwkomers) was soon drafted and passed through parliament without much discussion. This law, apart from laying down a number of conditions regarding the financial situation of prospective immigrants, requires them to have some very elementary ability in the Dutch language and some very basic knowledge of the Dutch society. These requirements are intended primarily as a minimal level of empowerment: to increase the chance for new immigrants to address people outside of the narrow circle that had been involved in organizing their immigration. The level of ability in Dutch and the degree of knowledge about Dutch society are both defined at a very low level, comparable to what serious tourists would wish to acquire when preparing a visit to a country which has their special interest. The assumption is that people who want to move in a new country and live there would have such an interest. Note the law does not address asylum seekers, but only people

[2] Any many other sites with similar titles: Russian women to marry, mail order brides, Russian ladies, Russia girls, Russian models, Ukraine love, Russian women videos, love tours, Moscow brides, order brides, Asian women.:

of foreign origin who supposedly out of free will decide they wish to settle in the Netherlands.

Once the law came to the implementation stage and the first reports on the test development were released, several parties outside of parliament started to voice concerns. Major opposition came from a university group with an interest in commercial applications of voice recognition software (e.g., train schedules and telephone number services) and from people who had some interest in the immigration, as prospective spouses, or as teachers to immigrants. The main concern of the latter group was that the requirements implemented in the tests would form a barrier to the immigration of people with limited means and or limited levels of education. One political party with 8 seats (out of 150) then choose to voice the concerns also in parliament. A number of journalists form newspapers and television engaged in the opposition, but were unable to find support for these views, not with the majority of the Dutch nor with recent immigrants they interviewed.

Having established in the preceding paragraphs that in spite of its potential for political controversy, the purpose and intention of the law on the integration for new residents were supported and considered justifiable by a vast majority both inside and outside of parliament the remainder of this article deals with the technical aspects of the test development.

The test system

Given the specific requirement of worldwide use and the additional requirement to avoid any dependency on level of literacy in the roman alphabet, it was decided to implement a test system using speech technology in the scoring process. The TGN uses Ordinate technology which provides an automatic scoring system that it is designed specifically for scoring the speech of language learners and can be administered over the telephone without any need for supporting written materials. Ordinate's technology was already operational for tests of English and Spanish and contains a number of language-independent components which can be used for any language for which the automatic scoring system has been trained. The scoring system was trained for Dutch using the speech of 1,500 learners of Dutch and of 800 Dutch native speakers.

What does the test measure?

The test measures 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. Listening skill is a prerequisite for all items. Thus, TGN tests both listening and speaking skills.

A1-minus and A2 levels

In principle the test measures over the range from no facility in Dutch up to and including a virtually perfect facility. The TGN is normed based on the levels of the CEF (Council or Europe, 2001), which at the request of the Ministry of Justice has been extended downward to include a level A1-minus (i.e., a level below A1), roughly corresponding to the level labeled 'Tourist' by North (2000). In keeping with the intended use of the test, measurement precision has been maximized at the lowest levels: from A1-minus up to and

including A2. On the basis of the recommendations of the Franssen Commission (Franssen et al., 2004a; 2004b) to the Minister of Immigration and Integration, the following level descriptors have been used:

> ***A1-minus***
> *Can communicate matters of direct personal importance using isolated words.* Uses isolated words, some standard expressions, and elementary polite expressions, but is difficult to understand because of pronunciation. Understands simple, everyday, concrete terms carefully pronounced and directed to him/her. Can also sometimes ask about such things using one or more isolated words. Conversation is not really possible.
>
> ***A2***
> *Communicates basic information about work, background, family, free time, etc.* Can make himself understood in short sentences, however pauses, false starts, and reformulations are prominently present. In general, pronunciation is clear enough to be understood despite a clear foreign accent. Uses a certain number of simple structures correctly but makes systematic elementary errors. Can connect word groups with simple conjunctions like "and", "but", and "because". Can understand clearly enunciating native speakers as long as he can ask for repetition when necessary.

Test administration conditions

Tests for immigration are administered abroad in the consulates and embassies of the Netherlands using landline telephones over a dedicated network of the Dutch Ministry of Foreign Affairs (MFA). Tests for naturalization are administered in the Netherlands in special testing centers. The same test is used in both settings. This article is limited to the application of the test abroad. Prior to the test, candidates get oral instructions in their native language or in another language that they say they understand sufficiently well. The instructions are given by a member of the embassy staff who is assigned as the test administrator and has received a special training. In cases where there is no suitable test administrator who can communicate well with the candidate, candidates themselves can bring someone to function as an interpreter for the oral instructions. Once candidates have understood the instructions, a telephone call is made to the computer that administers the test. After the automatic test delivery system has tested the acoustic quality of the telephone connection and finds it adequate, it begins to administer the test itself.

Candidates get 48 items of which 3 serve as warm-up/example at the start of each new section and 45 are used to determine the score. At the end of the test, there are two items in which candidates are asked to listen to a simple story and to retell it in their own words. The responses from these last two items are not scored. They are used for test validation and research purposes. The test, including the unscored items and the story retellings, but excluding the instructions, lasts about 12 minutes.

After the test is over, the results can be accessed a few minutes later via the Internet using the unique test identification number (TIN) of the candidate. When used from Dutch diplomatic posts to test prospective immigrants, test results are sent via e-mail to the administering post in the overseas countries and carbon copied to the Ministries of Justice and Foreign Affairs in the Netherlands.

Fraud

Given the high-stakes nature of the test precautions against fraud are of particular importance. The guidelines of the Ministry of Foreign Affairs anticipate measures to help prevent fraud in candidate identification and test administration. In addition, the TGN itself, to a certain extent, protects against potential fraud during administration and scoring. Obviously, the automatic scoring system makes fraud by exam raters impossible. Furthermore, during the administration of the TGN, each candidate gets a unique set of items. Items are randomly selected using stratified random sampling from an item bank. Through this procedure the probability that any two candidates sitting the test in a given location, within a given timeframe, have any items in common is negligible.

The item bank at launch in March 2005 contained around 1,000 items and is continuously replenished by means of item seeding. Fraud on the basis of foreknowledge of items is thus nearly excluded. Even when some items become disclosed, e.g., through memorization, the chance for candidates of getting those items on their tests is minimal and the impact on test performance will be negligible.

The Items

The TGN consists of three types of items.

Sentence Repeats

The candidate hears a spoken sentence and must repeat verbatim what was said. The sentence repeat items consist of a collection of sentences that occur frequently in spoken Dutch. They are drawn from authentic audio sources such as oral interactions and radio recordings. A large number consist of 'formulaic speech'. The stimuli are pronounced in an everyday, spontaneous manner as one would come across in normal spoken language use in the Netherlands. Stimuli vary in length from three to a maximum of thirteen words. The sentences are presented to the candidate in order of increasing difficulty. On short sentences, the candidate may be able to rely on short term memory, and ability to imitate pronunciation can play a role. On longer sentences, short term memory no longer suffices (Miller & Isard, 1956; 1963; 1964; Baddely, 1986; 2000) and the candidate must understand the sentence and utilize the sentence structure in order to reconstruct the sentence and repeat the stimulus word for word.

Examples are:

- *Ik heb gisteren naar die nieuwe TV-serie gekeken.* [I watched that new TV-series yesterday.]
- *Was nou eerst eens je handen*! [± Come on was your hands first!]

Short-answer questions

The candidate is presented with a short spoken question and must reply with a short spoken answer. This requires the ability to understand a spoken question and to respond with a relevant and comprehensible answer. Short answer questions ask for elementary information or simple conclusions with respect to time, quantity, lexical content, or logic. The requirement is that any Dutch native speaker should be able to answer the questions.

Or formulated alternatively, if the questions were asked in the native language of the candidates, they should have no problem to answer them. With a view to a potentially very low educational level of candidates in the target group, questions which require basic arithmetic skills were avoided. To be able to answer the questions, the candidate must be able to identify the words in the Dutch question, understand the words and their semantic relationships, interpret the question asked, formulate an answer in Dutch, and produce it in comprehensible Dutch.

Examples are:
Kun je rijst eten of drinken? [Can you eat or drink rice?]
Jan is ouder dan Piet. Wie is het jongst? [John is older than Pete. Who is the younger of the two?]

Opposites
The candidate must say the opposite of a given word. The words occur in everyday spoken language. Candidates must recognize and understand the word presented (passive vocabulary) and find the opposite (active vocabulary) and pronounce it. Ease in association of a word with its opposite gives an indication of passive and active vocabulary proficiency and is of importance in everyday conversation.

Examples are:
rechts [right]
donker [dark]

The Scoring Model

The automatic scoring system delivers four subscores. Two subscores measure *what* the candidate says (Vocabulary and Sentence skills) and two measure *how* the candidate says it (Pronunciation and Fluency). The four subscores are combined to form an Overall score. The minimum score is 10 and means that the candidate has not given any sign that he can understand and speak Dutch. The highest score is 80 and indicates that the candidate's Dutch language ability presents no impediment to communication with a native Dutch interlocutor about virtually anything. Items are recorded by native Dutch speakers. The speech may at times be lightly regionally colored, because to be representative, the speakers – women and men – were selected from different parts of the country.

Item Development

An initial set of over 3000 items was developed. All draft items were controlled for vocabulary against the Corpus of Spoken Dutch (CGN, 2004). This corpus contains recordings from telephone conversations, spontaneous conversations, interviews, and discussions in spoken Dutch, which could be used to determine which words were the most frequent, and from that, the relative importance of words.

All items were submitted in written form to experts in Dutch as a second language. All items that passed inspection were recorded by ten male and ten female speakers coming from different regions in the Netherlands. Test instructions were recorded by two

professional voice actors, a male voice for the general instructions for the exam and a female voice for the instructions for the different item types. All recordings were made in a professional sound studio.

In pretests administered over landline telephones, approximately 1,300 items were submitted to two target groups: 821 native speakers and 1,522 non-native speakers. The remaining items are reserved for item bank refreshment and will be pretested through seeding. Native speakers were included in the pretest to ensure that all items selected for live testing could easily be answered by native speakers, irrespective, of age gender and educational background. The pretests were delivered just like real tests: for each candidate the pretest consisted of a different selection of items. The pretests were delivered in immigrant schools (ROC's) in the Netherlands to recent immigrants. The students from the immigrant schools were complemented with subjects outside the schools, e.g., subjects inscribed in immigration programs, to achieve a suitable distribution of age and ability level.

The average age of the non-native speakers who took part in the pretest was 31 years. Subjects' ages ranged from 8 to 71. The ratio of men to women was 26:64 (N=1,341). Roughly 8% of the non-native speakers were illiterate. They came from 121 different countries. Roughly half had been in the Netherlands two years or less. One in five had not gone beyond elementary school. The average age of the native speakers was 37 years. Participants came from all parts of the Netherlands and had various educational backgrounds.

Based on the reactions of the pretest candidates, the items were quality controlled and answer models were determined. About 30% of the items were removed from the item pool because they failed to meet the previously defined quality criteria.

Development of the automatic scoring system

Apart from the purpose of trialing the items, the pretest dataset was collected for training the language-specific components of the automatic scoring system. The pretest speech data consisted of 132,000 utterances, of which about 59,000 were from native and 73,000 from non-native speakers. The data are quite varied: within the native speakers, many variants from different regions in the Netherlands, including dialects, are represented; within the non-native speakers, subjects coming from 121 countries worldwide are represented.

The total dataset is divided into subsets which are used to build initial scoring models, to train those models, and finally to validate those models. The goal of the automatic scoring system is the prediction of the language proficiency of the test taker as perceived by users of the language. The system is optimized to match human ratings.

Various human ratings were collected upon which to base the different subscores. To develop the content scoring (*what* the candidates say), trained native-speaker transcribers transcribed all the responses in the pretest data set. What the transcribers show that they can understand (because they can transcribe it) should also be what is understood by the automatic scoring system. Agreement between test scores based on the manual

transcriptions and the scores generated by machine is the ultimate criterion for the accuracy of the machine scores. For the development of manner scores (*how* the candidate speaks) human ratings of the candidates' responses on CEF-derived rating scales are collected for pronunciation and fluency. These serve to scale the machine scoring parameters.

In order to investigate the accuracy of the scores of the automatic scoring system, data from 139 subjects from the pretest were set aside for validation and not used for system development. These candidates' responses, which were not used in test development or scoring system development, were scored twice: once using the machine and once on the basis of transcriptions and ratings from specially trained human judges.

The Accuracy of the TGN

Reliability

Maximizing the accuracy of a test requires maximizing both the reliability and the validity of a test, in other words, maximizing the amount and the kind of information that a test provides. The quality and quantity of the items in a test and the degree to which they are tuned to the proficiency of the candidates together determine the amount of information that a test can deliver for a target group. It is desirable to maximize the amount of information near the decision boundaries. Indeed, to the extent that information is increased around a given point on the scale, the error around that point is reduced. Any test score on any test contains an amount of error. In classical test theory, reliability is defined mathematically as the ratio of the variance of the *true score* and the variance of the *observed score*. Modern test theory defines the standard error as the inverse of the square root of the information function, thereby recognizing that the standard error and the reliability, which depends on it, are not constant over the whole scale.

Individual estimates of the measurement error are available for all subjects who participated in the field test. These measurement error estimates are based on the difference on the theta scale between a subject's estimated ability and the difficulty of the items responded to by this subject. This is the conditional error of measurement. Its estimate is based on the sum of the information functions over all items presented to a subject. An important advantage of this index is that - unlike most reliability estimates – it is not dependent on the distribution of the subjects, but is determined locally on the theta scale. This index therefore provides a more direct estimate of the expected quality of measurements for subjects at different levels of ability. A complication caused by the random selection of items is that subjects are dealing with different subsets of items and that consequently no unique test information function can be computed. In fact when the item bank and the number of subjects become large there is an almost infinite set of information functions. Nevertheless it is possible to compute an estimated information function by averaging over a large set (several thousand) of possible item selections.

Table 1 shows a summary of the conditional standard error estimates at each of the lower bounds of the CEF cut-offs[3] on the TGN reported score scale. Two methods for computing these summaries were implemented: (1) by estimating the best fitting regression function

[3] The derivation of these cut-offs is explained later in this article.

(fourth order polynomial) and (2) by averaging over the estimated standard error for all subjects within an interval from three points below the cut-off to three points above it.

Table 1: *Conditional error of measurement on the TGN at the lower bounds of the CEF-levels estimated according to two methods (see text).*

CEF-level	Via regression	Averaging
A1-minus	2.92	2.82
A1	3.01	3.21
A2	3.20	3.14
B1	3.40	3.30
B2	3.62	3.49
C1	3.84	3.69
C2	4.04	3.95

Table 1 shows that both methods for estimating the conditional error of measurement yield comparable results. The error of measurement increases as the overall score goes up to the higher levels. Maximum precision is achieved at the lowest CEF level cut-off: A1-minus. Computing the reliability from these standard error estimates would yield a range of 0.961 – 0.972 at the A1-minus cut-off and a range of 0.941 – 0.963 for the A2 cut-off. At the cut-off for C1 the reliability range drops to 0.884 – 0.941and at C2 to 0.782 – 0.880.

Validity

In evaluating any device, it is important to select metrics that are relevant to the application for which it will be used and to exclude possible influence on these metrics from irrelevant factors. In evaluating the speech recognition technology implemented in the TGN functional precision is the most relevant criterion. Functional precision refers to the *effect* that specific errors in the system may have on the outcome of the application. In order to achieve functional precision it is not necessary for the system to correctly transcribe each and every word spoken by a candidate. For example, in a stock trading application, if the caller says “Sell 1000 shares of IBM Corporation at 72 good for the day, please” and the system recognizes “Sell 1000 IBM Incorporated at 72 good for the day” the correct action would occur. Therefore, functional accuracy is perfect even though the word error rate (WER) is 31%. However, if the system recognizes “Buy 1000 shares of IBM Corporation at 72 good for the day, please”, the functional accuracy is terrible even though the WER is only 8%. Functional accuracy ignores WER and instead looks at whether or not the right action is taken by the application. In the case of an automatic scoring system implemented in a spoken language test the action of the system is the assignment of scores to subjects based on their responses to a set of test tasks. This implies that functional precision is sufficient if the system assigns correct scores to the subjects.

Agreement with human ratings

Just as human transcribers, automatic speech recognizers make errors. Automatic speech recognizers even make more errors than well trained careful human transcribers (Lippman, 1996). But because the speech recognizer used in the TGN is one component in a complex of stochastically optimized components the speech recognizer must be evaluated against the effect of its errors on the reported TGN scores.

One way of evaluating the influence of the quality of the speech recognizer on the test scores is by scoring subjects' responses twice, once using speech recognition and a second time by rating human transcriptions of these responses. For this experiment we used the responses from 139 subjects which were not used in training any part of the system.

Figure 1 provides an overview of the experiment. The component labeled *ASR* stands for 'Augmented Speech Recognizer'. The component *Scoring logic* represents the application of the set of four measures (vocabulary, sentence skills, fluency and pronunciation) applied in the automatic scoring system. At the left hand side of Figure 1 subjects' responses to the three task types are shown: *Opposites*, *Short answers* en *Repeats*. For human based ratings all responses from the 139 subjects were manually transcribed. In addition responses to the Repeat tasks were rated by trained human raters in two separate rounds, one for pronunciation and one for fluency. Rating was carried out via telephone using Ordinate's testing system. All human 'ratings', i.e. the transcriptions and the pronunciation and fluency ratings subsequently passed through Scoring logic component resulting in a set of scores on the TGN reporting scale. The same responses were also fed into the ASR and the output of the ASR went through the Scoring logic. Finally correlations were computed between the sets of subscores and the Overall scores produced by human rating and those generated by automatic scoring. The correlation for the Overall scores was 0.93.

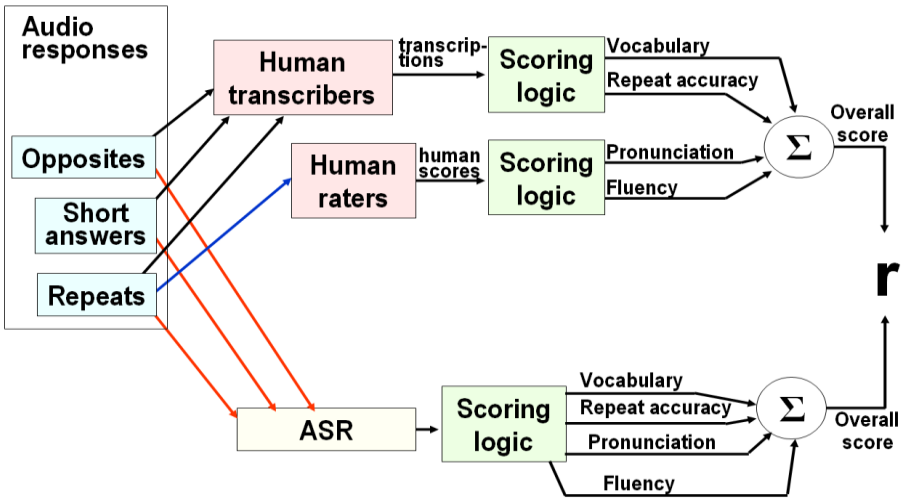

***Figure 1**: Assessing the functional accuracy of the SR in TGN*

The correlations for the subscores varied from 0.80 for pronunciation to 0.94 for sentence skills. Table 2 provides an overview of these correlations and the split-half reliability estimates for both sets of scores.

***Table 2**: Correlations between for ASR en human scores and reliability estimates (n=139)*

	Correlation	Reliability (split-half)	
	ASR~Human scores	ASR	Human scores

Pronunciation	0.80	0.89	0.94
Fluency	0.84	0.89	0.92
Vocabulary	0.85	0.73	0.78
Sentence skills	0.94	0.93	0.96
Overall score	0.93	0.94	0.96

The results presented in Table 2 show that the scores derived through the automatic system closely resemble those based on human transcriptions and ratings. Figure 2 shows a scatter plot of the human and machine generated scores. Subjects for whom the two types of scores differ significantly ($p<0.05$) are marked by a circle.

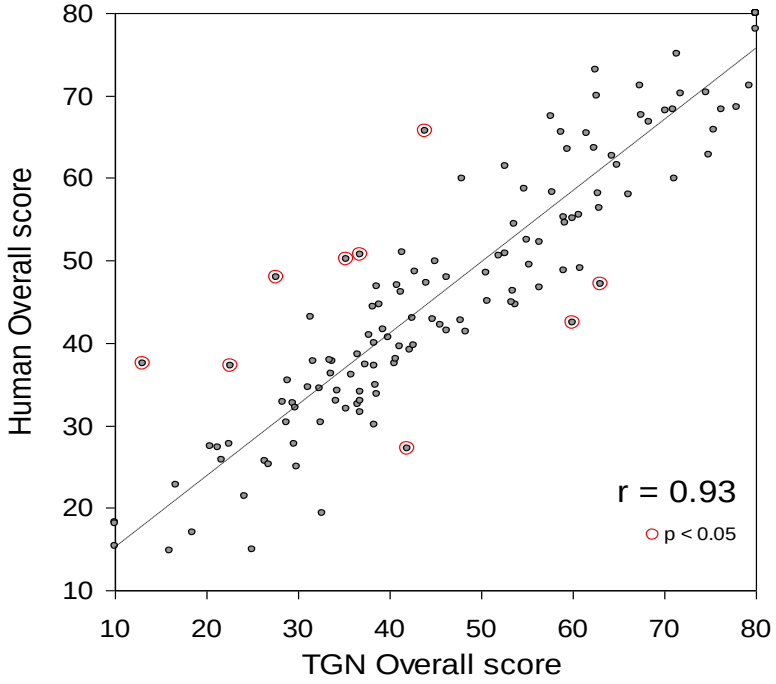

***Figure 2**: Scatter plot of TGN vs. Human Overall scores (n=139)*

Human ratings for Scaling

After having shown evidence for the functional accuracy of the automatic scoring system, evidence still needs to be collected to show these sores are predictive of the target ability. For this purpose candidates' oral skills were assessed using measures independent of the test scores. These data consisted of human judgments of candidates' oral skills. The judgments were from teachers as they experienced the candidates in class and from oral interview-based assessments using CEF scales, The data were gathered during the pretest and in three follow-up experiments. Table 3 provides an overview of the numbers of subjects in the experiments.

***Table 3**: Overview of non-native subjects involved in the development and validation of the TGN*

Experiment	**Number of subjects**	**Countries of origin**	**Subjects with spoken Dutch proficiency**	**Subjects with low educational attainment (max.**

			between 0 and A2	elementary school)
Pretest	1522	121	65%	20%
Follow-up experiment 1 (The Hague)	216	unknown	65%	unknown
Follow-up experiment 2 (Amsterdam)	353	57	90%	35 %
Follow-up experiment 3 (MFA/fit)	461	82	90%	25%

The subjects especially in the second and third follow-up experiments may be considered representative of the target group of the TGN as far as Dutch language proficiency and educational level are concerned.

Table 4 provides an overview of the main characteristics of all experiments involving human rating.

***Table 4**: Main characteristics of human rating experiments*

Study/ purpose	Material	Medium	Aspect	N (raters)	Rater ID's
Field test / test development	Story retelling	Telephone	Global	5	1-4, 8
Field test / validation	Classroom Observation	Life	Global	93	*53-145**
	Story retelling	Telephone	Global	5	1-4, 10
Den Haag / Phone quality	Classroom Observation	Life	Global		*147 – 213**
Amsterdam / validation	Structured interview	Life	Global, 1st rating	10	12, 14-15, 18, 21-23, 25, 27, 30
			Global, 2nd rating	10	12, 14-15, 18, 21-23, 25, 27, 30
	Career interview	Life	Global	13	11-13, 16, 18-20, 22, 26-27, 29-31
MFA-Fit / Item Fit & scaling	Structured interview	Life	Global, 1st rating	19	2, 32-49
			Global, 2nd rating	19	2, 32-49
		Recording	Global, 3rd rating	6	31, 34, 36, 39, 46-47
	Open questions	Telephone	Global, 1st rating	5	1-2, 50-52
			Global, 2nd rating	5	1-2, 50-52

* No or limited training with CEF

Evaluation of the human rating experiments

The set of ratings of the "story retelling' task was analyzed with version 3.54.1 of the computer program FACETS (Linacre, 1988; 2005). The reliability of the ratings on the CEF scale was estimated at 0.97. Out of the total number of paired ratings (1.743 cases) 68.2% were identical for both ratings in the pair.

In the Amsterdam experiment three ratings on the interview were gathered for 276 subjects. Ratings from the interviewer and the observer within the same interview showed 93% agreement and averaged at A1-minus. The rating from the separately conducted career interview also averaged at A1-minus, but exact agreement with the interviewer and observer in the other interview was respectively 51% and 48%. For a dichotomous decision (below or above the A1-minus cut-off) agreement between the interviewer and the observer was 98%, and each agreed with the career interview respectively 80% and 79%.

From the MFA-Fit experiment 243 complete cases were collected each with results from three tests and three human ratings on the interview. The ratings from the interviewer and the observer, each rating independently, agreed in 68% of these cases. Both ratings averaged at A1-minus. The agreement of these two ratings with the third independent rating based on the cassette recording was 40% and 41% respectively. For a dichotomous decision (below or above the A1-minus cut-off) agreement between interviewer and observer was, each agreeing with the third in 87% respectively 84% of the cases.

The ratings on the open questions gathered for the reliability study were analyzed using the computer program BEOVER (Heuvelmans, 2002). Because each question was designed to asses a different level, questions were analyzed separately. Table 5.2 provides an overview of the results.

Table 5*: Reliability study human rating of open questions (MFA-Fit)*

	Question 1	**Question 2**	**Question 3**
Estimate of variance components			
Subjects	82.8%	76.3%	81.4%
Raters	0.0%	0.5%	2.6%
Residual (error)	17.2%	23.2%	16.90%
Rater agreement	0.96	0.94	0.96
Estimated agreement for 2 raters	0.91	0.87	0.90

Based on the results presented in Table 5 it was decided there was no reason to exclude one or more of the raters for the main study. About 800 responses per question, 2402 in total were available for the main study. These ratings were used for scaling and standard setting described in the following paragraph.

Establishing the relation with the CEF-scale

The field test yielded insufficient data to establish the relation with the lowest CEF level (A1-minus) with comparable precision as could be achieved for the higher levels. In addition the field test contained a substantial proportion of human rated items that proved to be inadequate or yielded to little data. Because item context is known to be of influence on parameter estimation it was decided that additional data were needed for standard setting. This was one of the reasons for the additional experiments. Data from the experiments Amsterdam and MFA-Fit were collected with operational test forms and subjects were sampled to be representative of the lowest levels. A randomly selected subset from the field test data was combined with data from the MFA-Fit experiment to estimate the scaling functions and the relation with the CEF-levels. For validation of

these estimates a second non-overlapping subset from the field test was combined with the Amsterdam data.

CEF-ratings of the 'Story retelling' task from the field test were combined in a single data set with ratings of Open questions 2 and 3 and interview ratings from the MFA-Fit experiment and analyzed in a single FACETS-analysis. Open question 1 could not be combined with the two other questions because this question was scored dichotomously by the raters, 0 for below level A1-minus and 1 for A1-minus and above thereby attributing the same score to any subject above A1-minus, i.e., from A1-minus up to and including C2. Obtaining a score of 1 on this question would therefore represent a different value than on the other two questions which were scored continuously form 0 to 7 for all CEF-levels. The combined dataset contained a total of 5.984 ratings, involving 14 different raters, 1.009 subjects, performing three or more out of 60 different tasks. The link between the datasets was formed by raters that were involved in the both experiments.

Based on the FACETS analysis it can be concluded that the ratings were given with a high level of certainty. The reliability estimate for the total combined CEF rating data was 0.95. Out of the total of 3.777 paired ratings (two raters rating the same subject on the same task) 56% were exactly the same. This is 11% better than expected given the number of rating categories. Only 8 ratings had significant (p<1%) residuals. Inspection showed that these cases represented observed ratings at more than 1.5 levels distance from the expected rating. Figure 3 shows the probability curves for the CEF scoring categories. Each of the categories is well represented in the data and category curves cross the immediate adjacent curves at .5 indicating well established cut-offs for the CEF-levels.

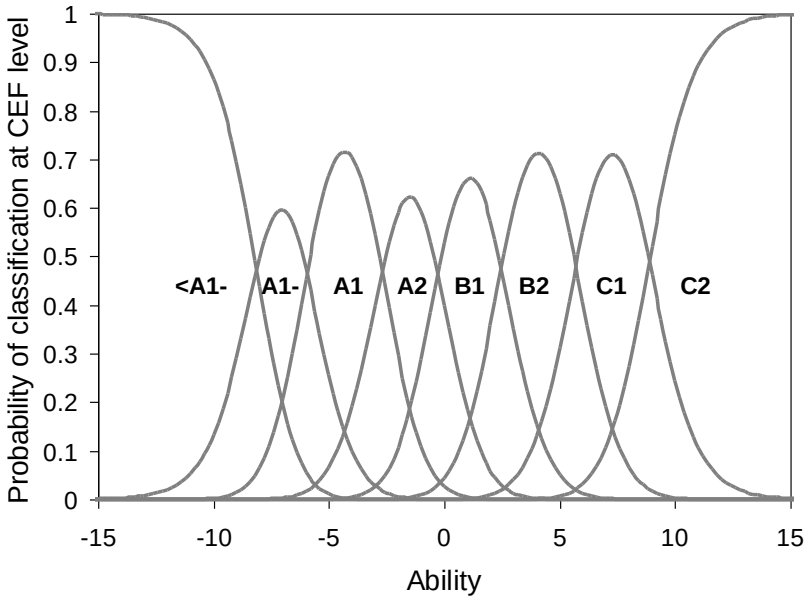

***Figure 3**: Probability curves for CEF score categories*

As can be expected given the size of the data no formal overall model fit can be established. Figure 4 however shows to what degree the data fit the model in practical

sense. De bolded curve from south-west to north-east represents the 'model', i.e., what level of ability (horizontal axis) is needed to obtain a rating in a particular scoring category (vertical axis). The dotted lines indicate the CEF-level cut-offs. The thin lines in parallel to the bolded curve indicate the 5% error level. Measurement error in an IRT context is expected to be larger at the extremes than toward the middle of the scale. In the present case it is clear that the error remains quite small at the lower end of the scale and only grows towards the upper end of the scale. This is according to expectation given the fact that both the subjects and the tasks were oversampled at the lower end of the scale. The grey dots represent average observations for a number of candidates at approximately the same level of ability[4]. The further these are located away from the bolded curve, the less well the data fit the model. The figure shows that the observations remain within the error bands indicating that the data fit the model in the practical sense.

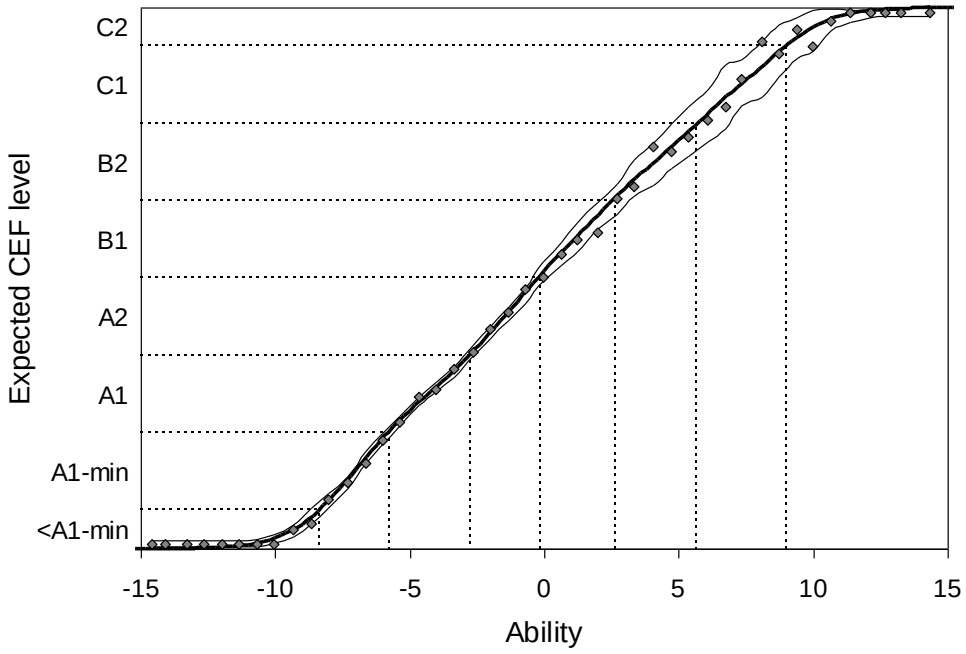

***Figure 4**: Estimated ability (modeled) and observations (data)*

The values of the estimates of the CEF cut-offs and their corresponding standard errors are reproduced in Table 6. Note that precision is higher at lower end of the scale. The estimates allow placing the subjects on the CEF scale and can then be used for setting the standard on the TGN by finding the correspondence between subjects' CEF rating and their TGN score.

***Table 6:** Estimated cut-offs for CEF levels based on human ratings*

CEF-level	Theta-CEF at lower bound	Standard error
A1-minuns	-8.13	0.07
A1	-5.96	0.05

[4] The FACETS programme forms as many groups as possible given the size of the dataset, each group containing about an equal number of subjects.

A2	-2.69	0.06
B1	-0.28	0.08
B2	2.47	0.15
C1	5.70	0.25
C2	8.88	0.27

Functions for transforming the underlying scales for the TGN subscales to the reporting scales of the subscores were established by the relationship of each of the theta scales of the subscores with the theta scale of the CEF. In a first step linear regression functions were estimated for projecting the CEF cut-offs on the underlying subscore scales. In a second step linear regression functions were used to transform these cut-off points on the underlying scales to a reporting scale. By using linear regression functions the interval properties of the theta scale are retained.

The theta scale with its values in three decimals running theoretically from minus infinity to plus infinity but with most of its estimated values in practice somewhere between – 6 and + 6 is usually rejected for reporting functions because of its lack of transparency for most score users. But in establishing the reporting scale one is free in choosing the values on that scale and can maintain the qualities of the original scale if values are transformed linearly. In educational systems scales are used with values from 1 to 5, from F to A, from 1 to 100 and many other variants. Most of these scales are delimited and do not have interval characteristics, though often people do as if they do by averaging over different narks obtained within the system. For TGN a scale with reported values between 10 and 80 is chosen. The lower and upper values have been chosen to indicate there are theoretically values outside of the reported interval, but they have no practical value. A score of 80 means an excellent performance but cannot guarantee 100% ‘perfect’ in all cases, whole a score of 10 intends to indicate that the subject has virtually no ability in the trait measured, though it cannot be excluded that she would be able to say a word or two. With the reporting scale running from 10 to 80 the seven cut-offs corresponding to the seven lower bounds of the CEF scale must be positioned at values between 15 for A1-minus and 80 for C2.

The overall score is to be computed by some combination of the subscores. In order to avoid that extreme subscores impact too heavily on the overall score, subscores are limited to the interval 0-90 after the transformation based on the regression functions. The four subscores are subsequently combined without further weighting to a total score (in fact differential weights have been estimated through the regression functions). Finally all subscores and the overall score are delimited to the interval 10-80 because in fact information outside that interval becomes irrelevant as the conditional standard error increases.

The estimated locations of the CEF lower bounds and their corresponding conditional errors of measurement on the TGN reporting scale are shown in Table 6.

***Table 6**: Relation of CEF-levels and TGN score*

CEF-levels	Reporting scale	Error
A1-minus	16	2.92
A1	26	3.01
A2	37	3.20
B1	47	3.40
B2	57	3.62
C1	68	3.84
C2	80	4.04

Further Validity Evidence

Validity evidence for the reported scores on a test can be obtained by investigating the relationships of these scores with external variables. For each external variable a hypothesis is formulated about the relationship. If the hypothesis cannot be rejected assumptions about the validity of the scores are supported. In the following paragraphs we will present a number of such hypotheses and the results of their testing.

Hypothesis: *Strong foreign accents cause Speech Recognition errors to increase, differentially penalizing candidates on content items, i.e., vocabulary and sentence skills.*
The hypothesis predicts that subjects will suffer more disadvantage from being scored by the automatic speech recognizer(ASR) if their pronunciation skills are relatively poor. Subscores for vocabulary and sentence level skills will be negatively influenced because the ASR has problems in dealing with their responses. To test this hypothesis we used the dataset with the 139 subjects from the field test who had not been involved in training the ASR. We evaluated the difference between their subscores for sentence level skills as derived from the ASR and from the human transcriptions as a function of their subscores for pronunciation based on human ratings. If the hypothesis were true, one would expect a positive correlation between the human scores on pronunciation and the difference between the ASR scores and the human scores on sentence skills. The result showed a small negative correlation of -0.20, whereby the hypothesis had to be rejected.

Hypothesis: *Items in a language test deal with content. World knowledge and experience will therefore have a positive influence on the scores obtained on a language test.*
The hypothesis predicts that generally speaking test scores will correlate with age. To test this hypothesis Figure 5 shows a scatter plot of TGN scores against the age of the subjects. Dark symbols represent native speakers and open dots represent the non-native speakers. In order to render more native speaker scores visible a randomly drawn value between -1 and +1 has been added to their scores if they were at the maximum of 80. Figure 5 clearly shows that native speakers obtained the maximum score, whatever their age and that non–native speakers score across the full range irrespective of their age. The hypothesis about a relation between age and score on the TGN must therefore be rejected. In addition Figure 5 shows that the TGN distinguishes very sharply between native speakers of Dutch and learners of the language.

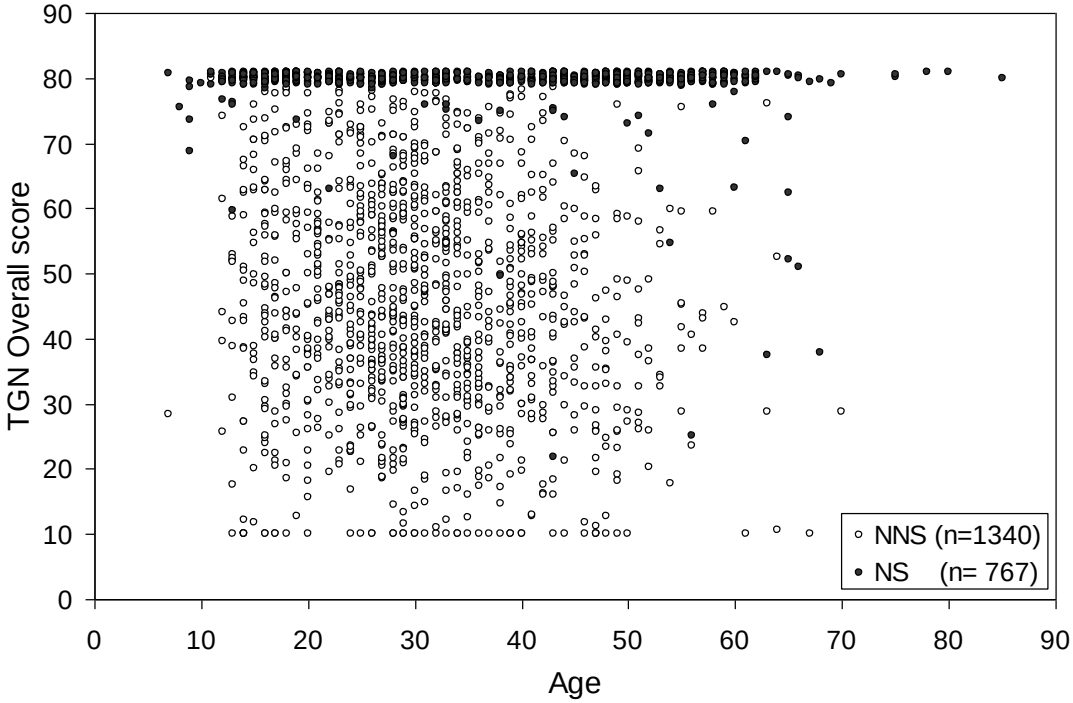

Figure 5*: Test scores in relation to age*

Several other hypotheses concerning unwanted relationships with the TGN, e.g., achieved level of education, gender, degree of literacy were tested and could be rejected on the basis of correlations close to zero, whereas relationships with variables such as years of residence in the Netherlands showed positive correlations (see full report in Dutch: Kerkhoff, Poelmans, De Jong and Lennig, 2005).

Accuracy of Pass-Fail Decisions

Finally, the accuracy of the TGN was further investigated through comparison between pass/fail decisions based on TGN scores and pass/fail decisions which would have been made based on scores from trained human raters rating candidates' performance in an OPI-like interview on a CEF-scale for oral interaction.. The TGN was developed in the first place to enable decisions about candidates seeking permission to immigrate to the Netherlands. An ability in spoken Dutch at level A1-minus was set as the required level. It is therefore relevant to evaluate to what degree the TGN corresponds to other measures evaluating candidates at level A1-minus. Table 7 shows the level of agreement between pairs of decisions. The paired results refer to the comparisons human-human, human-machine, and machine-machine. All results are based on the Amsterdam experiment and refer to the 228 subjects for whom complete data were available. In all pairs the ratings are independent and based on the results from the same sample of subjects. None of the data were used in training the system, in scaling or in standard setting.

The first two columns in Table 7 mention the source of the paired ratings. The following four columns provide the percentages of the 228 subjects for whom both ratings result in a pass at level A1-minus (1 – 1), the first results in a pass and the second in a fail (1 – 0) the first results in a fail an the second in a pass (0 – 1), and finally both result in a fail (0 - 0). The last column sums over the two columns where both ratings agree to pass the candidate (1 – 1) and to fail the candidate (0 – 0). The first two pairs involve human ratings only: the rating of the interviewer from the structured interview paired with the rating from the career interview and the rating from the observer at the structured interview with the rating from the career interview. So within each of these first two pairs subjects are rated in a different setting based on a different interview. For the third pair the maximum human rating (out of three available ratings) is paired with the maximum machine score (out of two available scores). The last pair involves the two available machine generated scores.

Table 7*: Agreement A1-minus cut-off: human-human, human-machine, and machine-machine*

Decision 1	**Decision 2**	**1 - 1**	**1 - 0**	**0 - 1**	**0 - 0**	**% Agree**
Interviewer	Career	64%	10%	9%	17%	80%
Observer	Career	63%	9%	12%	16%	79%
Human (max of 3)	TGN (max of 2)	70%	14%	7%	8%	78%
TGN-1	TGN-2	60%	7%	12%	21%	81%

From Table 7 it can be concluded that all pairs achieve approximately the same level of agreement, irrespective of the members of the pair. This means that well-trained human raters who base their CEF-rating on a structured interview or a career interview do not achieve better agreement than if their rating is compared to a score derived from the TGN and that the TGN score based on an automatic scoring procedure based on speech recognition data can predict a human rating as well as one human rating can predict another human rating. It must be noted, however, that the data do not allow us to decide which rating in any of the pairs is "right" when they do not agree: the true ability of the subjects remains unknown. On the other hand it can be maintained that two very different operationalizations – an automatic testing procedure and human CEF-ratings based on interviews – show as much overlapping decisions as two sets of human ratings do. It must therefore be concluded that both operationalizations to a large degree measure the same underlying variable and that there is a substantial overlap between the arguments on which human raters base their CEF-ratings and the set of subskills measured with the TGN.

Concluding remarks

Reliability estimates indicate that sufficiently reliable measurement can be realized using the TGN. The automatic speech recognition and scoring systems function sufficiently precise to generate ratings of subjects that are comparable with ratings from well-trained human raters. Within the collected data a relationship has been found between the test scores and relevant measures for mastery of Dutch spoken in interaction with Dutch speakers. Within the collected data no indication has been found that TGN scores are influenced by subjects' qualities and characteristics that can be assumed to have no relation with language ability.

The collected data furthermore indicate that after intensive training teachers using the CEF are quite able to provide reliable judgments on the language ability of subjects acquiring the Dutch language.

Based on the outcomes of the pretest and the follow-up experiments, the Dutch parliament agreed to the implementation of the TGN as a component in the decision procedures for admission of foreign subjects seeking to immigrate to the Netherlands for economic or sentimental (e.g., marriage, family reunion) reasons. However, the following condition was imposed: minimally the first 500 candidates were to be assessed by human raters in addition to the automatic scoring system. Human raters were to judge the candidates on the same items and applying the same criteria as used in the automatic scoring system. In addition the linking of the reported scores to the CEF-scale were to be re-evaluated. This research was to be performed by independent researchers from the Netherlands Organisation for Applied Scientific Research, known by its Dutch acronym: TNO. The results were reported by Kessens and Jacobusse (2007). Candidates were scored by four independent human raters. The automatic TGN score was found to correlate at 0.90 with the average over the four human raters. TNO concluded that there were no substantial differences between the automatic and the human scores. With respect to the linking to the CEF, TNO found that the distances between the adjacent levels (from A1- to A1, form A1 to A2, etc,) were correctly estimated, but TNO questioned the overall location of the TGN cut-offs with respect to the CEF scale, suggesting that TGN cut-offs may well have been projected at about one level too low. It has therefore been decided that further research is required. This research will involve a standard setting procedure and is to be carried out in 2008.

References

A.A. Aronowitz (2001). Smuggling and Trafficking in Human Beings: The Phenomenon, The Markets that Drive It and the Organisations that Promote It. *European Journal on Criminal Policy and Research, 9,2*, 163-195

Baddely, A. (1986) *Working memory*. Oxford: Clarendon Press.

Baddely, A. (2000) The episodic buffer: a new component of working memory? In: '*Trends in Cognitive Science'* 4(11), 417-423.

CBS (2003) Centraal Bureau voor Statistiek [National Statistics office]. Webmagazine 03 March.

CGN (2004). *Corpus Gesproken Nederlands*. Copyright (c) March 2004 Nederlandse Taalunie, Den Haag. Distrubutor: ELDA, Paris. S0113: Spoken Dutch Corpus.

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

De Jong, J.H.A.L. en M. Tijssen (2005). *Verantwoording Inburgeringexamen Kennis van de Nederlandse Samenleving.* [Validation report of the test Knowledge of Dutch Society]. Den Bosch: CINOP.

Franssen, J. et al. (2004a) *Inburgering getoetst. Advies over het niveau van het inburgeringsexamen in het buitenland.* Den Haag.

Franssen, J. et al. (2004b) *Normering inburgeringsexamen. Advies over het niveau van het nieuwe inburgeringsexamen in het Nederland.* Den Haag.

Heuvelmans, A. (2002*) OVERTON, A Computer Programme for Estimating Interrater Reliability*. Arnhem: CITO.

Hopkins , R. and J.Nijboer (2001) Country Report The Netherlands. In: Commission of the European Communities. Research based on case studies of victims of trafficking in human beings in 3 EU Member States, i.e. Belgium, Italy and The Netherlands. DG Justice & Home Affairs. Hippokrates JAI/2001/HIP/023

Kerkhoff, A., Poelmans, P., De Jong, J.H.A.L. en M. Lennig (2005) Verantwoording Toets Gesproken Nederlands.[Validation Report on the Test of Spoken Dutch] Den Bosch: CINOP.

Kessens, J.M. and G. Jacobusse (2007) Onderzoek naar de kwaliteit van het inburgeringexamen buiteland [Research into the quality of the immigration exam]. Utrecht: TNO.

Linacre, J.M (1988; 2005) *A Computer Program for the Analysis of Multi-Faceted Data.* Chicago, IL: Mesa Press.

Lippman, R.P. (1996) Speech perception by humans and machines. In: '*Proceedings of the European Speech Communication Association Tutorial and Research Workshop on the Auditory Basis of Speech Perception*', Keele University, UK, July 15-19.

Miller, G.A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. In: '*The Psychological Review'*, 63, pp. 81-97.

Miller, G.A. & Isard, S. (1963) Some perceptual consequences of linguistic rules. In: '*Verbal Learning and Verbal Behavior'* Vol. 2, 217-228.

Miller, G. A. & Isard, S. (1964) Free recall of self-embedded English sentences. In : '*Information & Control'* 7, 293-303.

North, B. (2000) *The development of a Common Framework Scale of Language Proficiency.* New York, NY: Peter Lang.

Odé, A.W.M. "002) *Ethnic-cultural and socio-economic integration in the Netherlands*. Assen: Van Gorcum.

Rouvoet Motion, TK 1998-1999, 25 403, no 30.

Van de Laan (2007). Confined Contact: Residential Segregation and Ethnic Bridges in the Netherlands. *Urban Studies, Vol. 44, No. 5-6*, 997-1017

Van Dijk, Jan (2002). Empowering Victims of Organised Crime: On the Concurrence of the Palermo Convention with the UN Declaration on Basic Principles of Justice for Victims. *International Review of Victimology, Vol. 9/1,* pp. 15-30.

Vocks, J. and J. Nijboer (2000). The Promised Land: A Study of Trafficking in Women from Central and Eastern Europe to the Netherlands. *European Journal on Criminal Policy and Research. 8, No 3,* 379-388.

APPENDIX: Item development for the Test "Knowledge of Dutch Society"

1 Preamble

In developing the item for the KNS a number of special factors had to be taken into account:

- The test is high stakes because candidates take the test in order to obtain a Temporary Permission for Residence in the Netherlands (*once admitted this can be turned into a permanent permission on certain conditions*).
- The test must be *administered worldwide in Dutch embassies and consulates (*non-professional staff).
- The target group is very diverse with respect to background s such as first language, level of education and ethnicity. The test must be accessible to illiterate candidates or candidates who have not mastered the roman alphabet.
- The test must be accessible to candidates at the A1-minus level as defined fo rthe test of Dutch language..

Apart form these, a number of additional principles were taken into account. The most important of these is that all 100 items are known and available in audio and in print as practice material. Candidates can prepare for the test by studying all questions. The questions must be answerable in Dutch at the A1-minus level. By answering the questions correctly candidates demonstrate that they have acquired some knowledge about Dutch society assuming that they have seen the film "Naar Nederland" (= to The Netherlands) in their own language and in Dutch.

Subject to be dealt with in the exam were inventoried by

- asking recent immigrants (up to 5 years of residence) what they would have liked to know before they came to live in the Netherlands;
- asking teachers of immigrants what they think would be useful knowledge for their students;
- a limited number, mainly in relation to history, was required by politicians.

Items were divided over 8 content areas:

1. Dutch geography, transport and living;
2. History;
3. State form, politics and laws;
4. The Dutch language and the importance to acquire it;
5. Upbringing and education;
6. Health care;
7. Work and income;
8. Taking the exam.

The exam consists of 30 questions. Ten different, equally difficult (Rasch equated), sets of 30 questions are compiled from the total "bank" of 100 questions. The questions are administered via telephone and are accompanied by 30 photographs (stills from the film) in an exam booklet. The 30 questions are the same as those as the candidate has been able to study in preparing for the exam.

2 Item construction

The principles have been operationalized as follows in the item development process:

2.1 Content of the items

- The items ask about knowledge facts that have been dealt with in the film 'Naar Nederland'.
- The items only ask about issues that have been clearly given attention in the film.
- Items refer directly to what has been told in the film and the same words as in the film are used.
- The items deal with all 7 parts of the film.
- The sequence or the items parallels the order of the film.

2.2. Accessibility for candidates at the A1-minus Dutch language level

Each question is phrased as simply as possible taking account of the A1-minus level.

- In the recorded question this is realized by a slow and clearly articulated speech, clear phrasing in meaningful units and realizing audible word boundaries and applying stress. For example:

  In Nederland / wonen daar / **véél** mensen / of / **weínig** mensen ?

  In the Netherlands / do there live / **many people / or / **few** people ?*

- The item msut be answerable with a single word or just a few words.
- It is assumed that candidates have practiced the 100 questions and answers in the photo book and the DVD with recorded answers and by practicing these with their teacher or via telephone with their partner or some other acquaintance in the Netherlands.
- The item must be visually supported by a still from the film.

2.3 Characteristics of the items

There are three types of items:

(1) Closed questions with a yes/no answer.

Question: Zijn de kranten, radio en televisie vrij in hun mening?
Are newspapers, radio and television free to express their opinion?

Answer: Ja
Yes

(2) Closed question with two options ("either…or" questions requiring repetition of the correct option)

Question: Kijk naar de foto. Wie is dit? Willem van Oranje of Prinses Maxima?
Look at the photo. Who is this? William of Orange or Princess Maxima?

Answer: Willem van Oranje.
William of Orange

(3) Open questions with a factual unambiguous answer.

Question: Wat zijn de kleuren van de Nederlandse vlag
What are the colours of the Dutch flag?

Answer: Rood, wit, blauw.
Red, white, blue

2.4 The vocabulary in the exam questions and answers

The size and the character of the productive vocabulary needed to answer the 100 KNS questions can be described as follows:

A total of 120 words is needed to answer all 100 questions. These are all tokens including proper names and numeric words.

Of thes 120 words:

- 35 are proper names that do not occur among the 500 most frequent words of the CGN(Corpus of spoken Dutch).,

But of those 120 words;

- 84 words occur on average 4.3 times in the questions;
- 103 words occur on average 22(!) times in the spoken text of the film 'Naar Nederland' thereby being considered as belonging to the category "personal experience" and "familiair";
- 61 words occur on average 30 times in the item bank for the language test.

2.5 Item difficulty

The figure below shows the relationship between the item difficulty (p-values) obtained from the total sample of field test subjects (n=on all field tested items (148) based on human (horizontal axis) and on machine (vertical axis) scoring. The figure also shows which item were retained (lighter dots) after the field test. Criteria for deletion were mainly high item difficulty and unacceptable discrepancy between human and machine scoring. The correlation between human and machine scoring went form 0.894 on all field test items to 0.948 on the retained items.

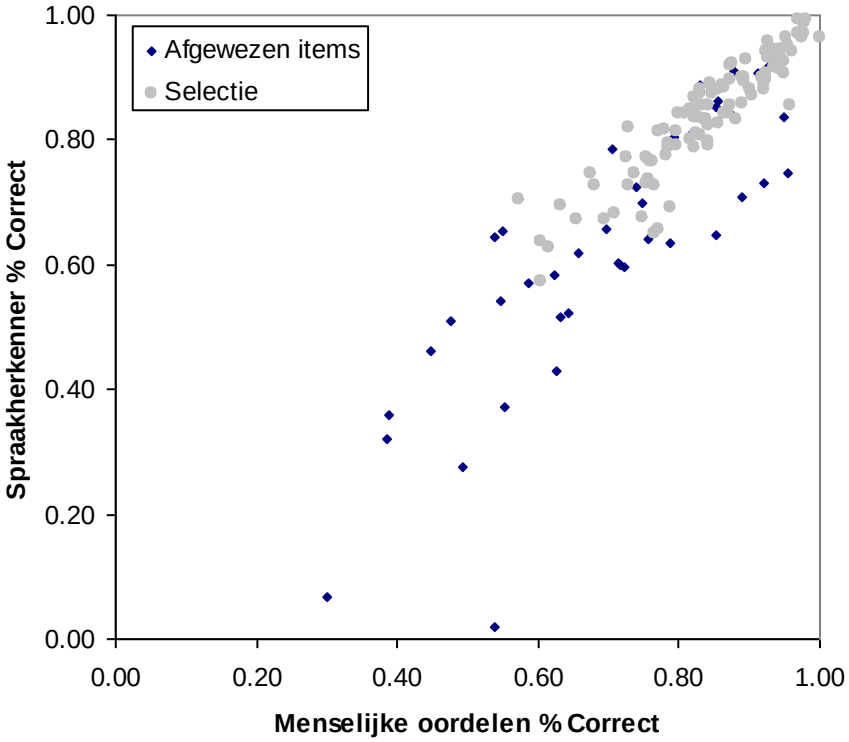

Table 3 presents an overview of the distribution of item difficulties (p-values) in the field test for 263 subjects at language level A1 or below. Roughly one quarter of all items obtained values of 0.9 or higher. About one half of all items obtained values between 0.8 and 0.9 and 92% of the items were answered correctly by one out of two of all field test subjects.

Table3: Proportion of correct answers according to human ratings

p-waarde	*CumFreq. %*	*CumFreq.Aantal*
≥ 0.90	24%	35
≥ 0.80	51%	75
≥ 0.70	73%	108
≥ 0.60	86%	127
≥ 0.50	92%	136
≥ 0.40	97%	144
≥ 0.30	100%	148
≥ 0.20	100%	148
≥ 0.10	100%	148
≥ 0.00	100%	148

Based on subjects at level A1 or below

Table 4 presents results for at the three lowest levels. The first column shows the percentage of corect answers (human ratings). Columns 2-4 shows the frequencies for the three levels (<A1-minus, A1-minus and A1), whereas columns 5-7 show the absolute numbers of subjects at thiose levels. It can be concluded that roughly one quarter of all subjects at the two lowest levels obtain an average of 90% correct or higher. . Almost half of all A1-minus subjects obtain a score of 80% ore above. Clearly a high score on the KNS can be obtained notwithstanding a low language level.

Table 4: Percentage of correct answers for subjects at the three lowest levels according to human ratings

% correct answers	*CumFreq. %*			*CumFreq.Number*		
	<A1-minus	*A1-minus*	*A1*	*<A1-minus*	*A1-minus*	*A1*
≥ 90%	24%	27%	54%	4	20	93
≥ 80%	29%	47%	66%	5	34	115
≥ 70%	41%	67%	75%	7	49	129
≥ 60%	41%	74%	84%	7	54	145
≥ 50%	47%	78%	91%	8	57	157
≥ 40%	76%	89%	92%	13	65	160
≥ 30%	88%	96%	95%	15	70	164
≥ 20%	94%	96%	97%	16	70	167
≥ 10%	100%	99%	97%	17	72	168
≥ 0%	100%	100%	100%	17	73	173