E8 Listening

Investigating the E8-Standards Listening Construct

LTC Technical Report 5

Bettina Wohlgemuth-Fekonja

Bundesinstitut
bifie

Alpen-Adria-Universität Klagenfurt
Universitätsstraße 64
AUSTRIA

Contents

1. The E8-Standards Listening Test

The E8-Standards Tests are taken by Austrian students of all ability groups in General Secondary School (APS) and Academic Secondary School (AHS) at the end of grade 8. They are diagnostic tests and as such report the students' strengths and weaknesses in the receptive skills of reading and listening and the productive skills of writing and speaking (Sigott et al., 2007:4-7). Their main objective is to evaluate the quality and efficiency of English teaching in Austrian schools, throughout all school types in eighth grade, and to consequently improve it on the basis of the results.

Against common concerns, the E8-Standards Tests do not request or force teachers of English to narrow down their teaching to a few aspects of the English language but to convey basic competencies and skills which students need in order to successfully communicate in the foreign language and to master their further education and/or professional training. One of the skills needed to absorb information, communicate with others, and exchange ideas and thoughts is the skill to listen.

The receptive skill of *Listening* was one of the two skills to be first introduced in the 3-year pilot phase of the E8-Standards Tests. Between 2006 and 2008 approximately 20,000 students were tested in *Listening*. In May 2009, 10 percent of all Austrian students in grade 8 participated in a Baseline Study, which forms the basis for future E8-Standards testing.[1]

E8-Standards Listening Test Specifications[2]

Generally speaking, test specifications define what a test tests and how it tests it. Test specifications play an important role in item production and moderation as well as in test design.

The E8 Listening Test Specifications in particular are a representation of the belief of what listening in English as a second language is and what

[1] http://www.uni-klu.ac.at/ltc/inhalt/520.htm (May 10, 2010)
[2] http://www.uni-klu.ac.at/ltc/downloads/LTC_Technical_Report_3.pdf (May 10, 2010)

Austrian students at the end of grade 8 have to have mastered to be considered proficient second language listeners.

As all specifications in language testing, the E8 Listening Test Specifications revolve around a construct. In the E8 Listening context, the construct is a clearly defined set of listening strategies, which students are supposed to be able to make use of when dealing with different sorts of topics and items. To be in the position to give differentiated feedback after the test, the items used in the E8 Listening Test are based on two principal areas of listening – *Direct Meaning Comprehension* and *Inferred Meaning Comprehension* – which are divided into sub–strategies (see Table 1).

Communicative Listening Strategies

1. Direct Meaning Comprehension

1.1. Listening for gist

1.2. Listening for main idea(s) or important information and distinguishing that from supporting detail or examples. This includes distinguishing fact from opinion when clearly marked.

1.3. Listening for specific information, including recall of important details. Understanding directions and instructions.

2. Inferred Meaning Comprehension

2.1. Making inferences and deductions based on information in the text. This can include deducing meaning of unfamiliar lexical items from context.

2.2. Determining a speaker's attitude or intention towards a listener or a topic

2.3. Relating utterances to their social and situational contexts

2.4. Recognising the communicative function of utterances

Table 1: Communicative Listening Strategies (Mewald et al., 2007:15)

E8-Standards Listening Item Characteristics

The above-mentioned strategies are part of the characteristics of the 143 E8 Listening items used in the pilot phase. Each item is based on an input text whose length determines the task design. A short task consists of a short input text (up to 100 words) and one item; a long task is comprised of a longer input text (between 200 to 500 words) and five items. An input text always revolves around one of seventeen different topics which, in the Austrian teaching and testing environment, are commonly referred to as the *17 Vertraute Themenbereiche*[3]. Text form, text type, domain, vocabulary and grammar are further characteristics defined for each task. In terms of the CEFR (Common European Framework of Reference) the difficulty of E8 items can either be below A2, A2 or above A2.

E8-Standards Listening Test Design and Administration

An E8 Listening Test is comprised of 20 4-option multiple-choice items, ten of which are based on short input texts. The remaining ten items are divided into two sets of five items, each of which is based on one longer input text.

An E8 Listening Test takes approximately 30 minutes. The students are presented with clear instructions at the beginning of the test, both by the test administrator and via the recording. Each recording is played to the students twice, including sufficient breaks in between. No questions must be asked during the test.

2. The E8-Standards Listening Test – A Valid Test?

Having been designed as a standardized test which predominantly fulfills diagnostic purposes and which will be conducted nationwide on a 3-year basis, it is indisputably necessary for the E8 Listening Test to be both reliable and valid. While reliability is concerned with the consistency of a set of measurements (Bachman et al., 1996:19), validity is concerned with the question of whether a test really measures what it is supposed to measure (Hughes, 2003:26).

[3] http://www.oesz.at/download/fss_hp/Kap2_Praxisreihe_9.pdf (December 15, 2009)

In the E8 context, the results generated in the course of the testing phase need to be trustworthy in order to be useful to the various groups interested in them, including students, parents, teachers, principals and school authorities, particularly because future teaching and school development will be based on these results. Hence reliability and validity are crucial in test development as well as in test administration.

“A reliable test score will be consistent across different characteristics of the testing situation” (Bachman et al., 1996:19) and generate similar results no matter when the test is taken. Administering and marking the tests consistently, providing clear test instructions and excluding ambiguous or faulty items contributes to the reliability of a test.

In addition to reliability, validity is one of the most important characteristics of a test. While a test might be perfectly reliable, it may at the same time not be valid at all, depending very much on the question of whether it really tests what it purports to test. If validity is missing, the scores obtained do not mean what they are supposed to mean. Consequently, decisions made on the basis of test results may go into a wrong direction and in this very specific context hardly serve any improvement in the school system.

Sireci (2007:477) states that “[validity] is not a property of a test [but rather] refers to the use of a test for a particular purpose.” He adds that to be able to make a clear statement about a test’s usefulness and appropriateness for a specific purpose one has to use several sources of evidence. This evidence, if sufficient, is then to be used to defend the purpose for which the test has been created and is administered. Sireci concludes by saying that “[evaluating] test validity is not a static, one-time event; it is a continuous process” (2007:477).

As Sireci points out, the evidence gathered to defend the purpose of a test needs to be ample and based on various types of validity. Before dealing with the validation of the E8-Standards Listening Test, the types of validity most

commonly referred to in language testing are to be briefly outlined in order to contextualize the approach to validation adopted in this study.

3. Validity and Validation

As language testing has been through various stages of development in the past century, test validation and types of validity have also undergone change. It is possible to distinguish between approaches which concentrate on test characteristics without taking into account any test scores, and those which are related to actual subjects' scores (Sigott, 2004: 44).

As already suggested, a test's validity can be examined in various ways, depending on the type of validity one is looking at. It is advisable to apply different methods to gather information about and evidence for whether a test is valid or not.

Alderson et al. (2005:171) name three main types of validity: *internal*, *external* and *construct* validity. While internal validity deals with "studies of the perceived content of the test and its perceived effect", external validity has to do with the comparison of subjects' scores with measures of their competencies taken from sources other than the actual test. Construct validity is the most complex type and sometimes seen as a general term for internal and external validity since it shares certain characteristics with both of them (Alderson et al., 2005:171f.).

Internal Validity

Face validity, *content* validity and *response* validity can be considered the most important types of internal validity. In the course of face validation, students and administrators – so-called non-testers – decide on a test's validity. Content validation is conducted by testers or subject experts who evaluate the test, and response validation is based on the interpretation of the test takers' feedback given in the form of self-report as well as self-observation (Alderson et al., 2005:172).

Face Validity

Within face validation non-experts, such as students or administrators, look at a test and determine whether the test's purpose is represented by its content (Alderson et al., 2005:172). In the early days of language testing, face validation was widely used by testers and primarily dealt with the question of whether a test looked as if it measured what it was said to measure. As language testing and test validation became more and more sophisticated, face validation was often strongly criticized and considered unscientific due to its lack of theoretical and scientific background (Sigott, 2004:45). According to Alderson et al. (2005:173) it nevertheless constitutes an important aspect of validation since test takers might perform totally differently, not to say better, on a test they consider to be valid and representative of the area they are tested in.

Content Validity

> Content validity is established by showing that the test items are a sample of a universe in which the investigator is interested. Content validity is ordinarily to be established deductively, by defining a universe of items and sampling systematically within this universe to establish the test (Cronbach et al., 1955:2).

While face validity relies on judgments gathered among non-experts, content validity is built upon a theoretical basis and has to do with judgments uttered by experts who comment on the test in a systematic way by first analyzing the content and then comparing that content with the test specifications set up before creating the actual test. In order to ensure a systematic approach to test validation, it is best to present the experts with either rating scales or "some precise indications of the aspects of the test" (Alderson et al., 2005:175), which the test's content can then be checked against. In contrast to face validation, where test developers appreciate the non-experts' feedback but might still not respect it, in content validation test developers are prepared to believe the experts even if they may disagree with their judgment (Alderson et al., 2005:173-175).

Anastasi (1988:131f.) names three things that need to be taken into consideration and adhered to when establishing content validity:

1. the behaviour domain to be tested must be systematically analysed to make certain that all major aspects are covered by the test items, and in the correct proportions;
2. the domain under consideration should be fully described in advance, rather than being defined after the test has been prepared;
3. content validity depends on the relevance of the individual's test responses to the behaviour area under consideration, rather than on the apparent relevance of item content.

Content validation has, however, one drawback. Disagreement among experts can lead to a surprisingly wide variety of judgments. If this is the case, test developers should go through other validation processes like external validation or face and response validation, and, if necessary, revise the test design as well as the specifications. Due to the risk of cloning, training judges to maximize agreement or getting judges together who usually agree is to be advised against. The panel of judges test developers consult should always be people whose feedback and conclusions they will respect and accept (Alderson et al., 2005:176).

Response Validity

Response validity is another aspect of internal validity. Response validation is based upon the test takers' response on test items. To obtain information on their behavior and thoughts while taking the test, most test developers gather introspective data retrospectively by having test takers explain to them why they chose the answers they did. The questions asked by the test developers should be open and not dictate the interview's direction. Retrospections may not always be very useful because test takers may not be able to recall why they chose a certain answer. This is where concurrent introspections in the form of think-aloud protocols could be used instead. While some automatic processes might not be able to be grasped, this sort of data collection might be very useful for tests in the course of which test takers are well aware of their processing – as in tests revolving around the productive skills of writing and speaking, for instance (Alderson et al., 2005:176f.).

External Validity

External validity can be divided into *concurrent* and *predictive* validity, which "may be considered together as *criterion-oriented* validation procedures" (Cronbach et al., 1955:1). The correlation coefficient is the statistic most commonly employed by test developers when carrying out external validation procedures.

Concurrent Validity

Concurrent validation deals with statistical rather than language-related information. While the different sorts of internal validation primarily focus on the test's content, concurrent validation concentrates on test scores. It is done by determining a criterion (Sigott, 2004:46) "which we believe is also an indicator of the ability being tested" (Bachman 1990:248) and by correlating this criterion with the test (Sigott, 2004:46).

> This other measure may be scores from a parallel version of the same test or from some other test; the candidates' self-assessments of their language abilities; or ratings of the candidate on relevant dimensions by teachers, subject specialists or other informants (Alderson et al., 2005:177).

A high correlation between test scores and criterion measure suggests that the test is valid (Sigott, 2004:46). For concurrent validation to be successful and meaningful, it is particularly important that the criterion, which the new test is checked against, is reliable and valid. Very often, however, a criterion whose validity has already been proven is not readily available. If test developers then decide to compare their experimental test with other tests that test takers have been presented with before but whose validity is still unknown, they need to bear in mind that the outcome needs to be interpreted with a lot of caution (Alderson et al., 2005:178).

Predictive Validity

In contrast to concurrent validation, in predictive validation the measures based on the criterion are not gathered at the same time as the test scores but after that.

> The simplest form of predictive validation is to give students a test, and then at some appropriate point in the future give them another test of the ability the initial test was intended to predict. […] A high correlation between the two scores would indicate a high degree of predictive validity for the [initial] test (Alderson et al., 2005:181).

Predictive validation is often used when examining the validity of proficiency tests (Alderson et al., 2005:180). Test developers could, for example, use a proficiency test meant to predict what a test taker's performance will look like when doing a graduate English course at a university. The criterion measure could either be his or her supervisor's evaluation of the test taker's ability to use the English language or the test taker's acquired knowledge at the end of the course (Hughes, 2003:29).

Sometimes the line drawn between concurrent and predictive validity can be very thin. Placement tests are, for example, also examined in terms of predictive validity. Test developers usually ask the teachers whether their students have been allocated to the right classes within the first week of teaching, before the students have had a chance to improve their skills. In this case the validation could either be seen as concurrent or as predictive validation (Alderson et al., 2005:182).

Construct Validity

A construct is the theoretical basis for what we think we are or would like to be measuring with a certain test. Before producing and administering a test, the construct underlying this test must be well defined and described. Consequently, the more clearly the construct is defined, the better the outcome of statistical analyses after testing will be (Weir, 2005:18).

"Construct validity refers to the degree to which inferences can legitimately be made from [a test] to the theoretical constructs on which [the test] was based" (Trochim, 2006). In order to determine the construct validity of a test, thorough empirical research as well as statistical evidence "[…] to support the inferences we make on the basis of test scores" (McNamara, 2003:467) is needed.

Ebel and Frisbie (1991:108) explain construct and construct validation as follows:

> The term construct refers to a psychological construct, a theoretical conceptualization about an aspect of human behaviour that cannot be measured or observed directly. Examples of constructs are intelligence, achievement motivation, anxiety, achievement, attitude, dominance, and reading comprehension. Construct validation is the process of gathering evidence to support the contention that a given test indeed measures the psychological construct the makers intend it to measure. The goal is to determine the meaning of scores from the test, to assure that the scores mean what we expect them to mean.

When undertaking construct validation, the theory underlying the construct is not questioned. Construct validation rather deals with the question of whether test developers have been able to operationalize the theory and create a valid test (Alderson et al., 2005:183). While within content validation one is solely interested in content coverage, construct validation combines the analysis of the content and the analysis of scores generated through a test (Sigott, 2004:47).

Based on Messick's (1989) thoughts on validity, language testers, when referring to construct validity, always also have to bear in mind the two threats to construct validity, namely 'construct under-representation' and 'construct-irrelevant variance' (cf. Brualdi 1999; Hamp-Lyons 1997 and Weir 2005). In case of construct under-representation, essential aspects or dimensions of the construct are not to be found in the test, which automatically leads to the conclusion that the test results do not provide the information about a test taker's ability the test should have yielded. Construct-irrelevant variance, on the other hand, refers to the reality that the data generated is based on various variables, which are of no relevance to the construct. Depending on the variables, the test might become easier or more difficult without testing what has originally been defined as the test's construct. If there is invalidity, a distinction between 'construct-irrelevant easiness' and 'construct-irrelevant difficulty' is to be made.

> "Construct-irrelevant easiness" occurs when extraneous clues in item or task formats permit some individuals to respond correctly or appropriately in ways that are irrelevant to the construct being assessed; "construct-irrelevant difficulty" occurs when extraneous aspects of the task make the task irrelevantly difficult for some individuals or groups. While the first type of construct irrelevant variance causes one to score higher than one would under normal circumstances, the latter causes a notably lower score (Brualdi, 1999:4).

As mentioned above, a test is validated by closely looking at the theoretical background underlying the test and by thoroughly interpreting the test scores. Construct validation encompasses different validation procedures and is therefore rich in information necessary to determine a nationally standardized test's validity, as needed within the E8 context.

> Construct validation studies can be seen as belonging to one of the following types: studies of test dimensionality, studies of sensitivity to treatments, studies of sensitivity to construct-external attributes, construct identification studies, and studies of mental processes (Sigott, 2004: 47).

4. Validating the E8-Standards Listening Test

Construct identification, in the course of which test content and method are related to test takers' scores, is the approach to be adopted in this report. By attempting to determine what makes test items difficult, construct identification leads to a deeper understanding of the test construct.

> The difficulty of any particular test, subtest or item depends on the kinds of abilities corresponding to the test content and method features that it taps. If tests, subtests or items differ in terms of the type or number of linguistic units that have to be processed, or in terms of the method features, and if these differences are correlated with the actual difficulty of the test, subtest or item, this can be interpreted as evidence that the tests, subtests or items indeed measure the abilities corresponding to the linguistic properties or the method features in terms of which the tests, subtests or items differ. Construct identification thus avoids the potential circularity inherent in other approaches to validation because the abilities of interest have to be made explicit before any data can be analysed (Sigott, 2004:52).

Objectives

The aim of this report is to gain more insight into, and develop a greater understanding of, the variables affecting the difficulty level of existing E8-Standards Listening Test items, which were developed by a pool of item writers, screened and commented on by experts, administered by BIFIE (Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens) and analysed by LTC (Language Testing Centre) at the University of Klagenfurt.

As mentioned before, the E8-Standards Listening Test serves a diagnostic purpose and is designed to assess the test takers' ability to listen to English as a second language.

Method and Results

The construct of the E8-Standards Listening Test is based on the assumption that listeners apply different strategies (*see Table 1: Communicative Listening Strategies*) depending on the item accompanying the aural input. Whether the strategies determine item difficulty or whether there are other factors influencing difficulty is to be investigated within the scope of the validation procedure applied in this context.

In the course of a 3-year pilot phase, lasting from 2005 to 2008, 143 E8 listening items were calibrated. Before piloting them, item writers and screeners decided upon an estimated level of difficulty, on the basis of which two different kinds of test booklets were put together. AHS- and 1st-ability-group HS-students were given a difficult version, whereas test takers belonging to second and third ability groups in HS were presented with an easier version.

Each test booklet comprised 20 items and showed an even distribution of items based on direct meaning comprehension and items based on inferred meaning comprehension. The different test booklets were linked with each other by using the same set of five pre-defined anchor items in every single test booklet. After a test and after analysis, the items were stored in the Klagenfurt Item Bank (KIB) to place the items on a common metric by means of dichotomous Rasch analysis.

The difficulty measures used for this report are the ones generated within the 3-year pilot phase. If an item was used more than once, an average

measure was computed. The difficulty measures range from -2.98 to 2.58, but were converted into a positive scale (0 – 5.56) to serve analysis purpose. For this analysis, the significance levels have been established as follows: highly significant ($p \leq 0.05$), slightly significant ($p > 0.05 \leq 0.1$) and not significant ($p > 0.1$). Item difficulty is the dependent variable. The independent variables chosen for analysis in the E8 listening context fall into three categories: item-related variables, text-related variables and item-text-related variables (cf. Grotjahn 2001).

Item-Related Variables

The variables of interest with regard to the items themselves are based on the stem, all four options together (including the solution), and the solution only.

Stem

1. *Number of words (V1)*: The number of words used in the stem, either formulated as a question or a sentence starter. It was assumed that the longer the stem, the more difficult the item.
2. *Number of negations (V2)*: The number of negations in the stem. It was assumed that the more negations in the stem, the more difficult the item.
3. *Stem type (V3)*: The stem was rated 0 if it was formulated as a question and 1 if it was formulated as a sentence starter. It was assumed that stems formulated as questions would be more difficult than stems presented as sentence starters.

Options

4. *Mean option length (V4)*: The average number of words in all four options. It was assumed that the longer the options, the more difficult the item.
5. *Number of negations (V5)*: The number of negations in the options expressed as a proportion related to the total number of words. It was assumed that the higher the proportion of negations in the options, the more difficult the item.

6. *Number of deictic words (V6)*: The number of deictic words, such as “he”, “there”, “those”, in the options expressed as a proportion related to the total number of words. It was assumed that the higher the proportion of deictic words in the options, the more difficult the item.

7. *Use of different tenses (V7)*: The options were examined in respect of whether the tenses used throughout all four options belonging to one item are the same or not. Options displaying only one tense were rated 0, options displaying different tenses were rated 1. It was assumed that the use of different tenses within a set of four options would increase item difficulty.

Solution

8. *Number of words (V8)*: The number of words used in the solution expressed as a proportion related to the total number of words used in all four options. It was assumed that the longer the solution, the more difficult the item.

9. *Number of deictic words (V9)*: The number of deictic words, such as “he”, “there”, “those”, in the solution expressed as a proportion related to the total number of words used in all four options. It was assumed that the more deictic words in the solution, the more difficult the item.

V1-9	N	Range	Minimum	Maximum	Mean		Std. Dev.	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
V1	143	14	1	15	6.99	.229	2.740	7.507
V2	143	14.29	.00	14.29	.5691	.21639	2.58761	6.696
V4	143	11.00	1.00	12.00	5.0944	.22076	2.63989	6.969
V5	143	25.0	.0	25.0	1.857	.3157	3.7751	14.251
V6	143	53.33	.00	53.33	5.6340	.76133	9.10412	82.885
V8	143	37.1	8.3	45.5	24.507	.4706	5.6275	31.669
V9	143	66.67	.00	66.67	6.6411	1.06298	12.71135	161.578

Table 2: Descriptive statistics for variables 1-2, 4-6 and 8-9

Table 2 displays the descriptive statistics for variables 1-2, 4-6 and 8-9. The correlations of variables 1-2, 4-6 and 8-9 with item difficulty are included in Table 3. Table 4 includes the results of two T-Tests for variables 3 and 7.

	Variable	*Pearson's r*	*p value (sig.)*
1.	Number of words in stem	.123	.143
2.	Number of negations in stem	.086	.307
4.	Mean option length	.282**	.001
5.	Number of negations in options	-.045	.596
6.	Number of deictic words in options	-.071	.397
8.	Number of words in solution	.109	.196
9.	Number of deictic words in solution	-.095	.257

Table 3: Correlations of variables 1-2, 4-6 and 8-9 with scale measure

	Variable	*Mean*	*t*	*p value (sig.)*
3.	Stem type		-.702	.484
	Stem type: Question (92)	*2.9118*		
	Stem type: Sentence starter (51)	*2.7693*		
7.	Use of different tenses in options		-.549	.584
	Use of different tenses in options: No (129)	*2.8026*		
	Use of different tenses in options: Yes (14)	*2.9821*		

Table 4: T-Tests for variables 3 and 7

Of all nine item-related variables only mean option length (4.) shows a slight relationship with item difficulty (Pearson's correlation coefficient = 0.282) (see Figure 1).

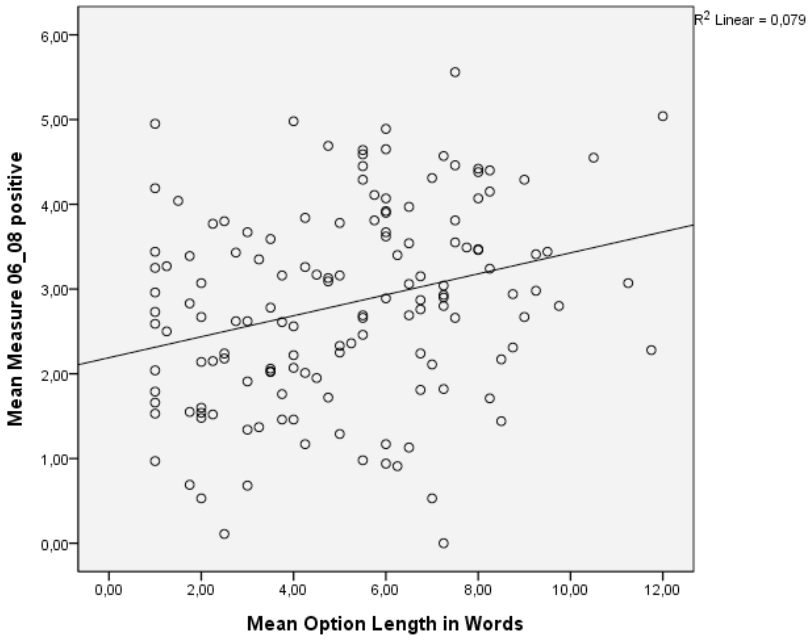

Figure 1: Scatterplot "Mean Option Length in Words"

Text-Related Variables

The variables looked at regarding the input texts themselves focus on text length as well as on the speakers' accent and the speed at which the texts are delivered.

10. *Accent (V10)*: The accents used in the recordings were divided into three categories: British English (1), American English (2) and Mixture of both (3). It was assumed that recordings displaying an American accent and recordings displaying a mixture of both, British and American accent, would be more difficult.
11. *Type of speech delivery (V11)*: Speech delivery was rated 1 if the input text was a monologue and 2 if it was a dialogue. It was assumed that items based on monologues were more difficult to solve.
12. *Number of words in input text (V12)*: The number of words used in the input text. It was assumed that the longer the text, the more difficult the item(s) belonging to it.
13. *Number of sentences in input text (V13)*: The number of sentences used in the input text. It was assumed that the more sentences in the text, the more difficult the item(s) belonging to it.
14. *Average sentence length (V14)*: The average number of words per sentence. It was assumed that the longer the sentences, the more difficult the item(s) belonging to the input text.
15. *Range of vocabulary in input text (V15):* The type token ratio computed for each input text. It was assumed that the higher the type token ratio, the more difficult the item(s) belonging to the input text.
16. *Length of recording in seconds (V16)*: The number of seconds a recording based on an input text lasts. It was assumed that the longer the recording, the more difficult the item(s) belonging to it.
17. *Average number of words per minute (average wpm) (V17)*: Speed was quantified by calculating the average number of words per minute. It was assumed that the more words per minute – the faster the speech delivery – the more difficult the item(s) belonging to the recording/input text.

V10-17	N	Range	Minimum	Maximum	Mean		Std. Dev.	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std.Error	Statistic	Statistic
V12	143	430	32	462	208.69	9.564	114.372	13080.947
V13	143	38	2	40	20.76	.898	10.739	115.323
V14	143	23.08	2.92	26.00	10.6099	.33733	4.03389	16.272
V15	143	1.00	1.07	2.07	1.6518	.02137	.25559	.065
V16	143	155	11	166	81.36	3.741	44.732	2000.992
V17	143	126.0	88.0	214.0	156.259	2.0545	24.5677	603.573

Table 5: Descriptive statistics for variables 12-17

The descriptive statistics for variables 12-17 are presented in Table 5. The results of a One-Way Anova analysis for variable 10 are included in Table 6. Table 7 displays the results of a T-Test for variable 11. The correlations of variables 12-17 with item difficulty are included in Table 8.

	Variable	*F value*	*p value (sig.)*
10.	Accent	.308	.736

	Variable	*Mean*
10.	Accent	
	Accent: British English (91)	*2.7722*
	Accent: Mixture (18)	*2.8067*
	Accent: American English (34)	*2.9556*

Table 6: One-Way Anova for variable 10

	Variable	*Mean*	*t*	*p value (sig.)*
11.	Type of speech delivery		1.843	.067
	Type of speech delivery: monologue (63)	*3.0198*		
	Type of speech delivery: dialogue (80)	*2.6629*		

Table 7: T-Test for variable 11

	Variable	*Pearson's r*	*p value (sig.)*
12.	Number of words in input text	.133	.113
13.	Number of sentences in input text	.043	.613
14.	Average sentence length	.085	.314
15.	Range of vocabulary in input text	.058	.491
16.	Length of recordings in seconds	.167*	.046
17.	Average number of words per minute (average wpm)	-.070	.404

Table 8: Correlations of variables 12-17 with scale measure

Of all eight text-related variables, type of speech delivery (11.) shows a slightly significant relationship with item difficulty (see Figure 2).

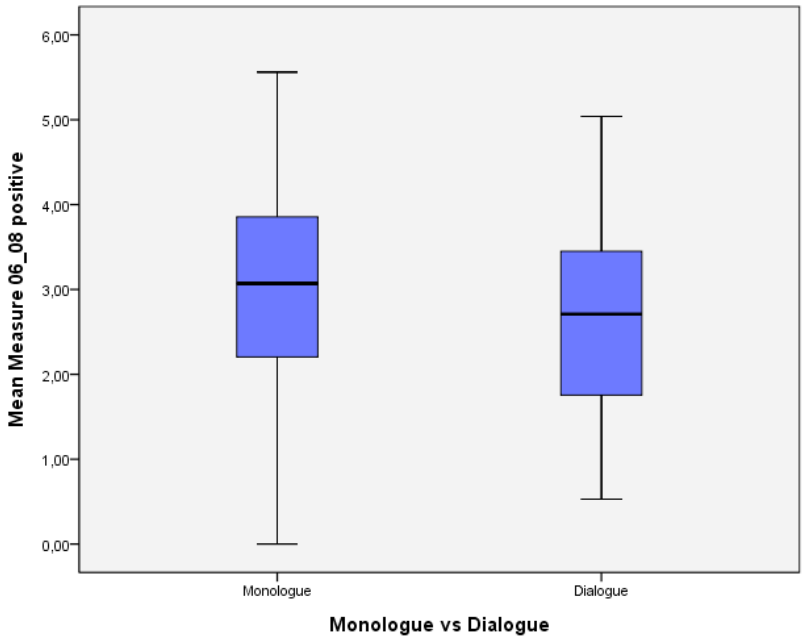

Figure 2: Boxplot “Monologue vs Dialogue”

Item-Text-Related Variables

Item-text-related variables deal with the relation between item and input text and mainly raise the question of whether the strategy one has to apply to solve the item, whether the linguistic overlap between item and input text and whether task format makes a difference in item difficulty.

18. *Strategy (V18)*: Each item is based on one of the listening strategies presented in Table 1. Which strategy an item is based on depends on the kind of problem-solving approach it is supposed to elicit. In the item development process each item was assigned a strategy. It was assumed that inferred meaning comprehension items were more difficult than direct meaning comprehension items.
19. *Item-text overlap (V19)*: Item-text overlap was divided into three categories: total overlap (2), partial overlap (1) and no overlap (0). Total overlap (2) corresponds to almost verbatim repetition of language material from the input text in the item. An item was rated 2 if at least two thirds of the

language used in it are actually used in the input text. Partial overlap (1) applies to items which include synonyms, antonyms, hypernyms and hyponyms of words used in the input text. No overlap (0) refers to items, which neither include any of the content words nor any synonyms, antonyms, hypernyms or hyponyms of words mentioned in the input text, which predominantly applies to indirect meaning comprehension items. It was assumed that item difficulty would decrease with the amount of overlap given – the more overlap, the easier the item.

20. *Task format (V20)*: Each item is based on a task. Short tasks consist of a shorter input text and one item; long tasks consist of a longer input text and five items belonging to the text. It was assumed that items based on long tasks were more difficult than items based on short tasks.

The results of two One-Way Anova analyses for variables 18 and 19 are included in Table 9 and Table 10. Table 11 displays the results of a T-Test for variable 20.

Variable	*F value*	*p value (sig.)*
18. Item Strategy	4,084	,001

Variable		*Subset for alpha = 0.1*	
18. Item Strategy	*N*	*1*	*2*
2.2. Determining a speaker's attitude or intention towards a listener or a topic	9	2,3489	
1.2. Listening for main idea(s) or important information and distinguishing that from supporting detail or examples. This includes distinguishing fact from opinion when clearly marked.	21	2,4486	
2.3. Relating utterances to their social and situational contexts	13	2,5031	
1.3. Listening for specific information, including recall of important details. Understanding directions and instructions.	45	2,5631	
1.1. Listening for gist	12	2,7150	
2.4. Recognizing the communicative function of utterances	12		3,3600
2.1. Making inferences and deductions based on information in the text. This can include deducing meaning of unfamiliar lexical items from context.	31		3,5465
Sig.		,418	,637

Table 9: One-Way Anova for variable 18

	Variable	*F value*	*p value (sig.)*
19.	Item-Text Overlap	8,410	,000

	Variable		*Subset for alpha = 0.1*	
19.	Item-Text Overlap	*N*	*1*	*2*
	Total overlap	34	2.1718	
	Partial overlap	69		2.9264
	No overlap	40		3.1880
Sig.			1.000	.271

Table 10: One-Way Anova for variable 19

	Variable	*Mean*	*t*	*p value (sig).*
20.	Task format		-1.980	.050
	Task format: short (53)	*2.5725*		
	Task format: long (90)	*2.9660*		

Table 11: T-Test for variable 20

All three variables have a highly significant influence on item difficulty. With regard to item strategy it was found that, based on a 0.1 significance level, the seven different strategies fall into two subsets of different difficulty. Data additionally show an overall effect of strategy on item difficulty (see Figure 3). Concerning item-text overlap, the categories partial and no overlap form a group, as opposed to total overlap, which constitutes a group of its own (see Figure 4). With regard to task format, items belonging to long tasks are tendentially more difficult than items belonging to short tasks (see Figure 5).

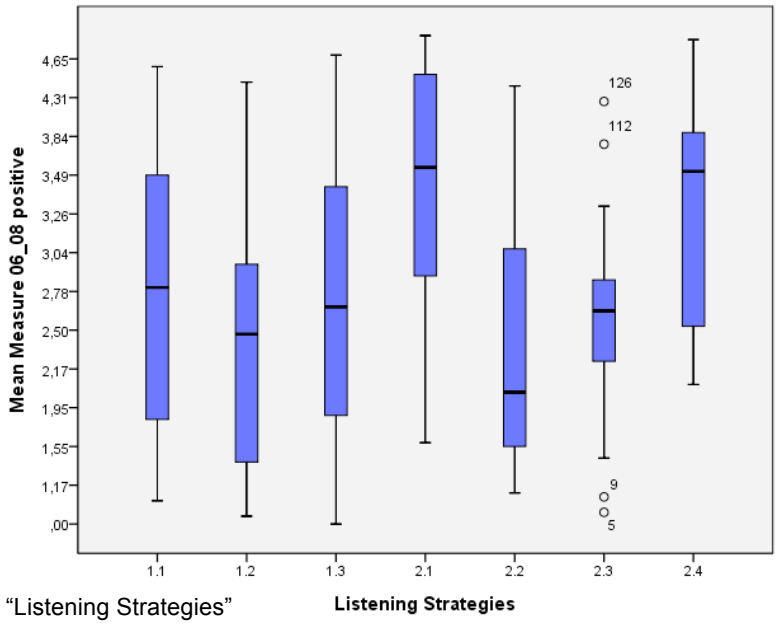

Figure 3: Boxplot “Listening Strategies”

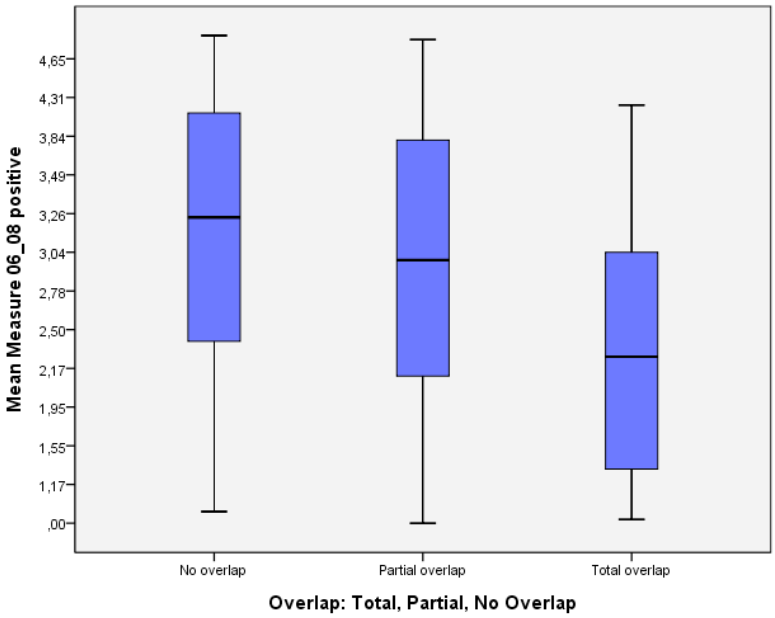

Figure 4: Boxplot “Overlap”

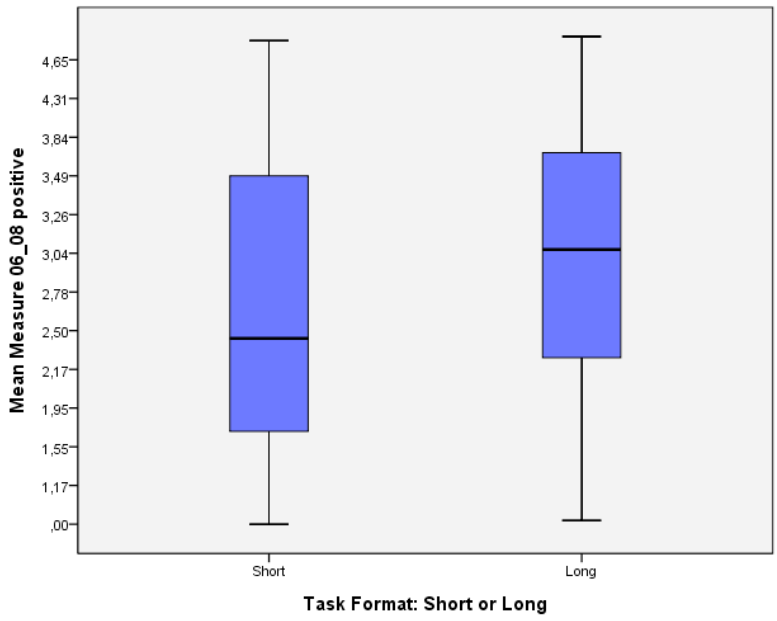

Figure 5: Boxplot “Task Format”

5. Conclusion

The aim of this report was to investigate the E8 Listening construct and, by doing so, to gather information on variables affecting item difficulty. The variables were divided into three categories – item-related, text-related and item-text-related variables – each being further divided into sub-variables which, in the preliminary stages, had been identified as variables worth investigating.

Out of the twenty variables studied, five have a significant effect on item difficulty. Among the item-related variables, mean option length has a significant influence on item difficulty, meaning that as the number of words in the options increases, so does item difficulty, which makes sense considering that processing more words in the options asks for more language comprehension. It also suggests that memory may play a vital role in processing longer options.

Among the text-related variables, the type of speech delivery has an effect on item difficulty. Items based on monologues are generally more difficult than items based on dialogues, which could be explained by the different discourse structure. Monologues are frequently similar to written texts, while dialogues are usually characterized by features such as repetition, back-channeling, reformulation, repair, or turn-taking, which can aid comprehension.

Out of the three item-text-related variables, the strategies underlying the items are the first variable that significantly correlates with item difficulty. As already mentioned, analysis has shown that the seven strategies used in the E8 context fall into two separate categories, proving that certain strategies clustering in one category or sub-group do not extensively differ in item difficulty among each other but are of different difficulty from one category or sub-group to the other. The classification of certain strategies into various sub-groups of different difficulty has other implications too, namely that language learners acquire those strategies that are easier for them to apply earlier than the ones which are still difficult for them to make use of.

Concerning item-text overlap, partial and no overlap fall into one group, suggesting a clear difference between items displaying almost verbatim repetition of the language material used in both, the input text and the item, and items which are either characterized by only partial or even no overlap. Items consisting of almost the exact words used in the input text are clearly easier to

solve than items including synonyms, antonyms, hypernyms or hyponyms of words used in the input text or items which do not show any word-related connection with the input text at all.

The third item-text-related variable that has been subjected to an analysis is task format, distinguishing between short and long tasks. Analysis has shown that items belonging to long tasks are significantly more difficult than items belonging to short tasks, implying that the more information testees have to process, the more difficult the items because of both, the cognitive load and the longer attention span necessary to successfully solve the item.

Having chosen construct identification as an approach to validate the E8 Listening Test, variables, which might be potential indicators of item difficulty, had to be defined. Depending on their characteristics, each of the twenty variables was assigned to one of the three main groups relevant for analysis: item-related, text-related and item-text-related variables. Out of these three groups, the item-text-related variables have turned out to be the most crucial variables when discussing differences in item difficulty. Item-text overlap, task format and the listening strategies underlying the items make for the range in difficulty.

Having collected evidence of the validity of the E8 Listening Test, test developers can now use the information gathered within the validation process for future endeavors, both in further validation studies as well as in well-guided item production.

6. List of Tables and Figures

7. Bibliography

Alderson, J.C., C. Clapham and D. Wall. 2005. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Anastasi, A. 1988. *Psychological testing*. New York, NY: Macmillan.

Bachmann, L.F. and A.S. Palmer. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

Brualdi, A. 1999. *Traditional and Modern Concepts of Validity*. ERIC/AE Digest. ED435714.

Ebel, R.L. and D.A. Frisbie. 1991. *Essentials of Educational Measurement*. Englewood Cliffs. New Jersey: Prentice-Hall.

Grotjahn, R. 2001. "Determinants of the Difficulty of Foreign Language Reading and Listening Comprehension Tasks: Predicting Task Difficulty in Language Tests." In H. Pürschel and U. Raatz (Eds.): *Tests and translation*. *Papers in memory of Christine Klein-Braley*. Bochum: AKS 2001, pp. 79-102.

Hamp-Lyons, L. 1997. "Washback, impact and validity: ethical concerns." *Language Testing*, Vol.14, No.3, pp. 295-303.

Hughes, A. 2003. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

McNamara, T. 2003. "Book Review: Fundamental considerations in language testing. Oxford: Oxford University Press, Language Testing in practice: designing and developing useful language tests." *Language Testing*, Vol.20, No.4, pp. 466-473.

Messick, S. 1989. “Validity”. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103), New York, NY: The American Council on Education and Macmillan.

Mewald, C., O. Gassner and G. Sigott. 2007. *Testing Listening. Specifications for the E8-Standards Listening Tests. LTC Technical Report 3*. Language Testing Centre, Alpen-Adria-Universität Klagenfurt.

Sigott, G. 2004. Towards Identifying the C-Test Construct. In: Language Testing and Evaluation (editors R. Grotjahn and G. Sigott), Frankfurt am Main: Peter Lang.

Sigott, G., O. Gassner, C. Mewald and K. Siller. 2007. *E8-Standardstests. Entwicklung der Tests für die rezeptiven Fertigkeiten: Überblick. LTC Technical Report 1*. Language Testing Centre, Alpen-Adria-Universität Klagenfurt.

Sireci, S.G. 2007. “On Validity Theory and Test Validation.” *Educational Researcher*, Vol.36, No.8, pp. 477-481.

Weir, C.J. 2005. *Language Testing and Validation – An Evidence-Based Approach*. Hampshire: Palgrave Macmillan.

Internet

Cronbach, L.J. and P.E. Meehl. 1955. “Construct Validity in Psychological Tests.” *Psychological Bulletin*, 52, pp. 281-302.
http://psychclassics.yorku.ca/Cronbach/construct.htm (accessed 2 December 2009)

Trochim, W.M.K. 2006. *Idea of Construct Validity*.
http://www.socialresearchmethods.net/kb/considea.php (accessed 28 December 2009)